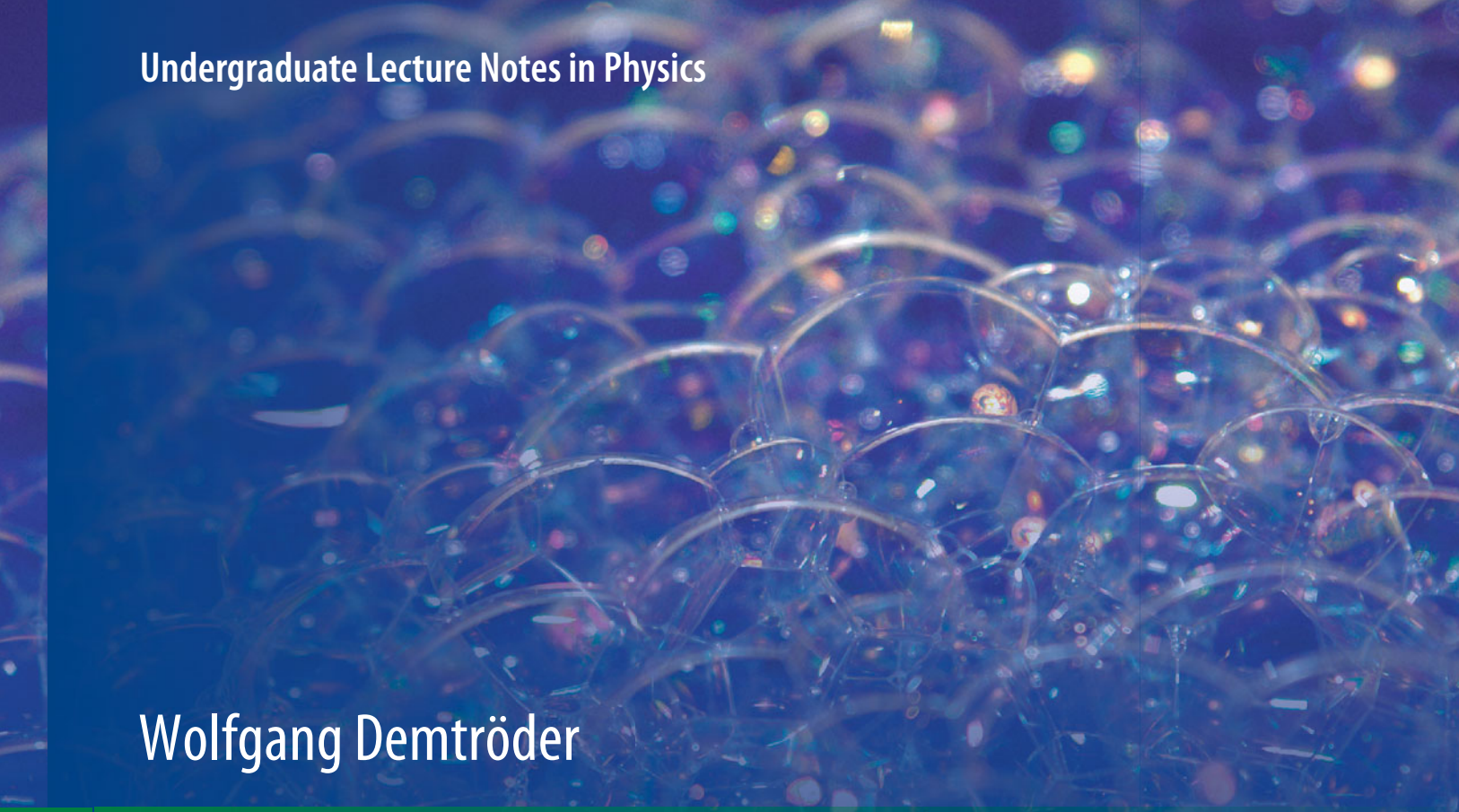
Undergraduate Lecture Notes in Physics



Wolfgang Demtröder

Electrodynamics and Optics

Springer

Undergraduate Lecture Notes in Physics

Undergraduate Lecture Notes in Physics (ULNP) publishes authoritative texts covering topics throughout pure and applied physics. Each title in the series is suitable as a basis for undergraduate instruction, typically containing practice problems, worked examples, chapter summaries, and suggestions for further reading.

ULNP titles must provide at least one of the following:

- An exceptionally clear and concise treatment of a standard undergraduate subject.
- A solid undergraduate-level introduction to a graduate, advanced, or non-standard subject.
- A novel perspective or an unusual approach to teaching a subject.

ULNP especially encourages new, original, and idiosyncratic approaches to physics teaching at the undergraduate level.

The purpose of ULNP is to provide intriguing, absorbing books that will continue to be the reader's preferred reference throughout their academic career.

Series Editors

More information about this series at http://www.springer.com/series/8917

Wolfgang Demtröder

Electrodynamics and Optics

Springer

Wolfgang Demtröder
Universität Kaiserslautern
D67663 Kaiserslautern
Rheinland-Pfalz, Germany
e-mail: demtroed@rhrk.uni-kl.de

ISSN 2192-4791 ISSN 2192-4805 (electronic)
Undergraduate Lecture Notes in Physics
ISBN 978-3-030-02289-1 ISBN 978-3-030-02291-4 (eBook)
https://doi.org/10.1007/978-3-030-02291-4

Library of Congress Control Number: 2018958365

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Contents

Electrostatics

1

Electrostatics treats the phenomena caused by static electric charges. Nearly 2000 years ago the Greek already gained first qualitative experiences with electric effects. They discovered that amber (greek: *electron*) became electrically charged by rubbing.

Today we have, besides a more fundamental knowledge, also a great number of technical applications of electrostatics. A few of them will be discussed in detail. Nevertheless, many basic questions have still to be answered. They are based on a deeper fundamental knowledge of atomic physics and we refer to volumes three and four of this textbook series for more details.

1.1 Electric Charges; Coulomb's Law

During the last three hundred years, many experimental investigations revealed the following facts [1, 2]:

(a) There exist two different types of charges: positive +, and negative − charges. They can be distinguished by the forces between each other and also by the deflection of charged particles in electric and magnetic fields (see Sects. 1.8.2 and 3.3).
(b) Charges with equal signs repel each other while charges with opposite signs attract each other (Fig. 1.1). Contrary to the force of gravitation where only attraction exists, we now have repulsive as well as attractive forces. These forces can be used to measure the amount and the sign of electric charges.
(c) Charges are always associated with particles with mass. The most important carriers of negative electric charges are the electron and negative ions. Negative ions are atoms or molecules with excess electrons.

The most important carriers of positive electric charges are the atomic nuclei and positive ions. Positive ions are atoms or molecules that miss at least one electron. Furthermore, there exist short living elementary particles with negative or positive charges, i.e. pion, muon, positron and anti-proton.

> Charges with opposite signs attract each other while charges with equal signs repel each other.

The positive charge $+e$ of the proton and the negative charge $-e$ of the electron represent the smallest values of electric charges found hitherto.

All charges Q in nature are an integer multiple of this elementary charge e. Exceptions are the quarks, supposed components of hadrons (heavy particles, see Vol. 1, Sect. 1.4) with charges of $1/3 \cdot e$ respectively $2/3 \cdot e$. However, according to our knowledge today, quarks do not exist as free particles.

Very accurate measurements have shown that the amounts of the charges of proton and electron differ at most by $10^{-20}e$ and we have good arguments to suppose that they are exactly equal (see Vol. 3).

The total charge inside a closed system remains temporarily constant, i.e. electric charges cannot be created or destroyed.

Note, however, that charges of one sign can be isolated by spatially separating positive and negative charges (see

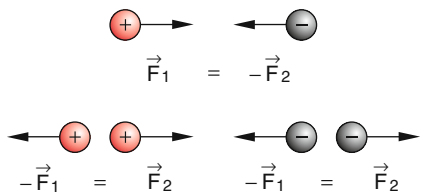


Fig. 1.1 Attraction of charges with opposite sign and repulsion between charges with equal sign

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_1

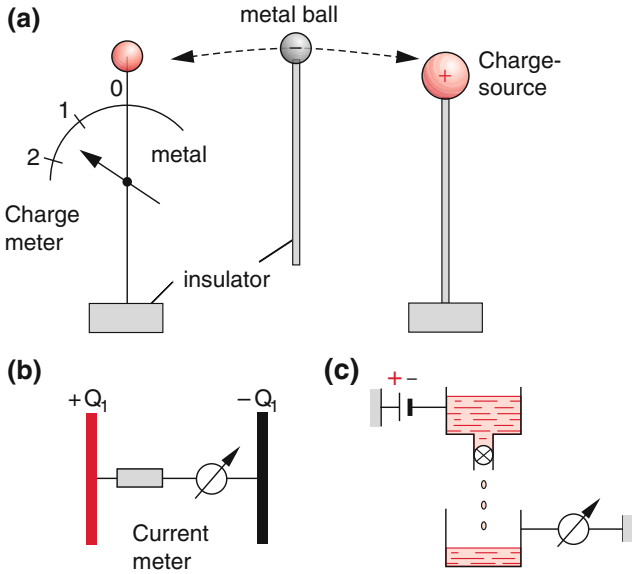


Fig. 1.2 Transport of charged particles **a**) with a "spoon", a metallic ball on an isolating rod, **b**) via an electric conductor between the charged metallic plates, **c**) by charged droplets of water

Sect. 1.5). An example is the ionization of hydrogen atoms where electron and proton are separated.

- Charges can be transported between a source of charges and a measuring instrument by an electrically isolated metal ball (see Fig. 1.2a) but also by electric conductors (Fig. 1.2b) or charged droplets of water (Fig. 1.2c).

The transport of electric charges represents an electric current. Transport of charges is always connected with the transport of mass.

Note Because our environment is electrically neutral, charges of one sign are "created" by spatial separation of negative from positive charges, where the condition must be fulfilled that the algebraic sum of positive and negative charges is always zero.

Examples

Triboelectricity (separation of charges by friction), emission of electrons from the surface of a hot cathode, ionization of atoms.

The forces between two charges Q_1 and Q_2 and their dependence on their mutual distance r can be measured quantitatively with Coulomb's torsion balance (Fig. 1.3). This is a device analogue to the Eotvos' balance for measuring the gravity force. (see Vol. 1, Sect. 2.9).

An electrically isolated rod is suspended by a thin wire. At a distance L from the axis a metallic ball is mounted at

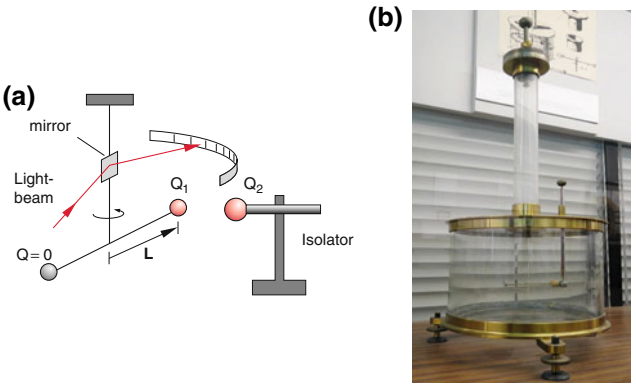


Fig. 1.3 Coulomb balance **a**) Schematic arrangement **b**) original construction as reproduction in the Deutsches Museum

one end of the rod and a counterweight at the other end. Now the ball is charged and another charged ball is moved towards it. The force between both charges generates a torque $\boldsymbol{D} = \boldsymbol{L} \times \boldsymbol{F}$ which is balanced by the opposite torque exerted by the twisted wire. Measuring the twist angle for different distances r between the charged balls yields the important equation for the force between charges Q_1 and Q_2, named (Fig. 1.4) **Coulomb's law** (Fig. 1.3)

$$\boldsymbol{F} = f \cdot \frac{Q_1 \cdot Q_2}{r^2} \hat{\boldsymbol{r}}, \tag{1.1}$$

with a proportionality constant $f > 0$, and the unit vector $\hat{\boldsymbol{r}}$ in the direction from Q_2 to Q_1 (Fig. 1.5). Equation (1.1) shows

Fig. 1.4 Charles Augustin de Coulomb (1736–1806)

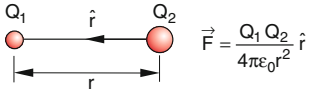


Fig. 1.5 Forces between charges

that the force $\boldsymbol{F}$ and the vector $\hat{\boldsymbol{r}}$ are parallel for charges of equal—repulsive force—and anti-parallel for charges with opposite sign—attraction.

In Eq. (1.1) the units for the force $\boldsymbol{F}$ (Newton, N) and the distance r (meter, m) are fixed. So only the units for the proportionality constant f and the charge Q can be chosen.

1.1.1 Systems of Measurement

During the historical development of physics mainly two systems of units evolved. These are the earlier CGS (cm, gramm, sec) system and the SI (systeme internationale). The CGS system is still preferred by theoretical physicists. In this book we will only use the SI.

1.1.1.1 The International System of Units (SI)

The SI has already been introduced in Vol. 1, Sect. 1.7. The unit of electric charge Q is derived from the electric current I, i.e. the amount of charges transported through the cross section A in one direction during the unit of time (second, s).

The current I is the fourth basic unit of the SI, (ampere, A) and can be expressed by the units of length and force (see Sect. 3.3.2 and Vol. 1 Sect. 1.6.8). So the unit of charge is

$$[Q] = \text{Coulomb} = \text{C} = \text{A s}.$$

Coulomb's experiment yields for the force between two equal charges of 10^{-4} C and 1 m apart

$$F = f \cdot \frac{10^{-8}\,\text{C}^2}{1\,\text{m}^2} = 89.875\ \text{N}.$$

The constant f in (1.1) becomes $f = 8.9875 \times 10^9\,\text{Nm}^2/\text{C}^2$. As will become evident the constant f is chosen as

$$f = \frac{1}{4\pi\varepsilon_0}$$

and includes the dielectric constant ε_0 of the vacuum

$$\varepsilon_0 = 8.854 \times 10^{-12}\,\text{A}^2\text{s}^4\text{kg}^{-1}\text{m}^{-3},$$

Coulomb's law is written in SI-units as

$$\boldsymbol{F} = \frac{1}{4\pi\varepsilon_0}\frac{Q_1 \cdot Q_2}{r^2}\hat{\boldsymbol{r}}. \tag{1.2}$$

The unit of ε_0 can be written as *As/(Vm)* *because* $1\,\text{kg}\,\text{m}^2\,\text{s}^{-2} = 1\,\text{Nm} = 1\,\text{VA}$. The unit Volt, (V) is explained in Sect. 1.3.1.

Examples

1. An electron carries a charge of $-e = -1.6 \times 10^{-19}$ C.
2. If it were possible to remove one electron of each atom in two equal masses of 1 kg copper, containing about $N = 10^{25}$ atoms this would represent a positive charge of $+\text{N} \cdot \text{e} = 1.6 \times 10^6$ C. The force between these two bodies separated by 1 m would be $|\boldsymbol{F}| = 2.3 \times 10^{22}$N!!

1.1.1.2 The cgs System

The cgs system sets the factor f in Coulomb's law equal to the dimensionless number one. The force is measured in dynes, the length in centimeters (cm). We then get from $[F] = [Q^2/r^2]$ the unit of charges

$$[Q] = [r] \cdot [F]^{1/2} = 1\ \text{cm} \cdot \text{dyn}^{1/2}\ \text{named ESU (electrostatic unit)}.$$

$$1\ \text{ESU} = 1\ \text{cm} \cdot \sqrt{\text{dyn}}$$

The cgs system is often used in theoretical physics because of fixing the constant f to $f = 1$ simplifies the notation of many equations. But the great disadvantage is that you have to know every conversion factor between mechanical and electromagnetic units.

Therefore, in this textbook we will use solely the internationally agreed units of the SI.

The SI unit Coulomb expressed in ESU is

$$1\,\text{C} = 3 \times 10^9\,\text{ESU}$$

1.1.1.3 Measurement of Charges

Charges can be measured with an electrometer. Figure 1.6a shows the realization with a metallic pointer Z that can rotate about the axis D.) It takes the equilibrium position when the acting torques $\boldsymbol{D}_g$—caused by gravity and $\boldsymbol{D}_c$—caused by electrostatic forces between the pointer and the vertical support just cancel each other.

The center of mass of the pointer lies below the axis D. Therefore without charges the pointer shows zero at the top of the scale. Charges brought to the instrument create a repulsive Coulomb force between the metallic pointer and the metallic vertical support resulting in a deflection of the pointer because they carry charges of the same sign.

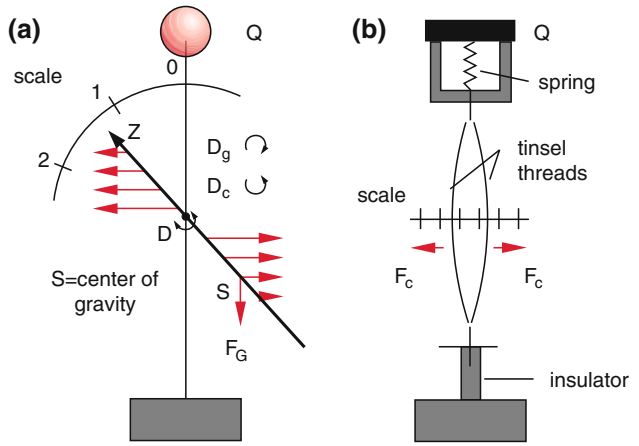


Fig. 1.6 Measurement of charges

The electrometer with two filaments shown in Fig. 1.6b uses the repulsive force between thin Lametta strings to measure the amount of charges.

Note With both instruments only the amount of charges can be measured but not the sign.

Charges can not only be measured by their forces on each other but also by discharging a capacitor through a conductor of high resistance. The time variation of the current $I(t)$ is measured (Fig. 1.2b).

For the total charge we get

$$Q = \int_0^\infty I(t)\,\mathrm{d}t.$$

Note Equation (1.1) for the force between two charges is mathematically equivalent to the law of gravitation. The ratio of gravitational and Coulomb's force is

$$\frac{F_G}{F_C} = \frac{G\dfrac{m_1 \cdot m_2}{r^2}}{\dfrac{Q_1 \cdot Q_2}{4\pi\varepsilon_0 r^2}} = 4\pi\varepsilon_0 \cdot G \cdot \frac{m_1 \cdot m_2}{Q_1 \cdot Q_2}.$$

Examples

1. We consider two balls of lead each of mass $m = 10$ kg which carry a charge $Q = 10^{-6}$ C. At a distance of 0.2 m between their centers the Coulomb force is $F_C = 0.22$ N but the gravity force is only $F_G = 1.7 \times 10^{-7}$ N. Their ratio is then $F_G/F_C = 7.7 \times 10^{-7}$.
2. We replace the balls of experiment 1 by two electrons of mass $m_e = 9.1 \times 10^{-31}$ kg, and charge $Q = -e = -1.6 \times 10^{-19}$ C. This gives the ratio

$$\frac{F_G}{F_C} = \frac{4\pi\varepsilon_0 G \cdot m^2}{e^2} = 2.4 \times 10^{-43}!$$

3. Electron and proton of the hydrogen atom attract each other at a distance of 0.5 Å $= 5 \times 10^{-11}$ m with the Coulomb force of $F_C = 9.2 \times 10^{-8}$ N. The corresponding gravitational force is smaller by a factor of 4.4×10^{-40}.
4. The repelling force between two protons in the atomic nucleus with a mean distance of $r = 3 \times 10^{-15}$ m is $F_C = 26$ N. Because the nuclei are stable there must be a compensating attractive force (nuclear force). Since the force of gravitation between these protons is only 2.1×10^{-35} N gravitation cannot be responsible for the stability.

These examples illustrate that gravitational forces in micro physics can be completely neglected against the Coulomb forces.

The strong electrostatic forces are responsible for the relative high energy demanded to separate charges in macroscopic bodies. The following example will illustrate this fact.

Consider a ball of radius 1.5 cm made of electrically neutral copper. If it were possible to ionize only 1% of the 1.2×10^{24} atoms and transfer these electrons onto an equal but neutral ball placed 1 m apart then each body would bear an excess charge of $\Delta Q = \pm 1.9 \times 10^3$ C. The attractive force between the balls would be 3.3×10^{16} N.

Macroscopic bodies are in general electrical neutral. Therefore Coulomb's forces of positive and negative charges cancel each other. Then the gravitational forces become dominant.

Even in the microscopic region (attraction or repulsion between two atoms) the electrical forces between two neutral atoms nearly outweigh. However, the positive and negative charges have different spatial distributions and therefore the Coulomb forces do not cancel completely (see Sect. 1.4.3).

The chemical bond does not only depend on the Coulomb interactions which can be attractive but also repulsive. In addition exchange interactions play an important role which can be explained only by quantum mechanics.

1.2 The Electrostatic Field

In Vol. 1, Sect. 2.7.5 we had introduced the gravitational field G that is independent of a test mass m. A more common concept is the electric and magnetic field. Here we also have to find a field which is independent of the test charge.

1.2.1 Electric Field-Strength

The force

$$\boldsymbol{F}(\boldsymbol{r}) = \frac{Q \cdot q}{4\pi\varepsilon_0 r^2}\hat{\boldsymbol{r}}, \tag{1.3}$$

of a charge Q located at the origin of a coordinate system acting onto a test charge q at the position $\boldsymbol{r}$ can be measured. The charge Q creates a force $\boldsymbol{F}(\boldsymbol{r})$ according to (1.3) with a magnitude still dependent on the test charge q. To overcome this dependence we define an electric field $\boldsymbol{E}(\boldsymbol{r})$ as the quotient $\boldsymbol{F}(\boldsymbol{r})/q$ of force and test charge

$$\boldsymbol{E}(\boldsymbol{r}) = \frac{Q}{4\pi\varepsilon_0 r^2}\hat{\boldsymbol{r}}. \tag{1.4}$$

We name this expression the *electric field strength* of the charge Q and the corresponding normalized force field $\boldsymbol{F}(\boldsymbol{r})/q$ the electric field. Its unit is

$$[E] = [F/q] = 1\,\text{N/A s} = 1\,\text{V/m}.$$

$$\boldsymbol{F} = q \cdot \boldsymbol{E}. \tag{1.5}$$

If there are several charges Q_i distributed in space then we find the total force upon a charge q by vector addition of the individual forces (Fig. 1.7). Now we position the test charge q at the origin and the field charges Q_i at the positions $\boldsymbol{r}_i$.

The total force is then

$$\boldsymbol{F} = \frac{q}{4\pi\varepsilon_0}\sum_i \frac{Q_i}{r_i^2}\hat{\boldsymbol{r}}_i. \tag{1.6a}$$

The total field strength at the position of the test charge q is then $\boldsymbol{E} = \boldsymbol{F}/q$.

Besides point-charges there exist quasi continuous charge distributions with the spatial charge density $\varrho(\boldsymbol{r})$, where ϱ is defined as the charge per unit volume (Fig. 1.8). The total charge in volume V is

$$Q = \int_V \varrho(\boldsymbol{r})\,\mathrm{d}V.$$

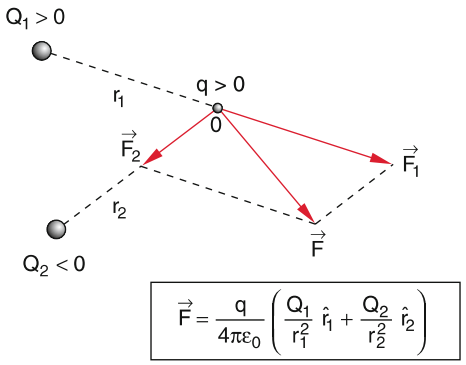


Fig. 1.7 Force on a test charge q by different field charges Q_i

where we have chosen the origin arbitrarily.

The force $\boldsymbol{F}(\boldsymbol{R})$ upon a test charge q at a point $P(\boldsymbol{R})$ outside of the volume V with the space charges $\mathrm{d}Q = \varrho\,\mathrm{d}V$ is, according to Fig. 1.8

$$\boldsymbol{F}(\boldsymbol{R}) = \frac{q}{4\pi\varepsilon_0}\frac{\boldsymbol{R}-\boldsymbol{r}}{|\boldsymbol{R}-\boldsymbol{r}|^3}\varrho\,\mathrm{d}V. \tag{1.6b}$$

The force between the total charge Q and the test charge q is then

$$\boldsymbol{F}(\boldsymbol{R}) = \frac{q}{4\pi\varepsilon_0}\int_V \frac{\boldsymbol{R}-\boldsymbol{r}}{|\boldsymbol{R}-\boldsymbol{r}|^3}\varrho(\boldsymbol{r})\,\mathrm{d}V. \tag{1.6c}$$

Correspondingly we treat electrically charged surfaces, e.g. metal plates which have a surface charge density $\sigma = Q/A$. The total charge of the area A becomes

$$Q = \int_A \sigma\,\mathrm{d}A \tag{1.6d}$$

Generally speaking we can conclude:

The presence of point charges Q_i or of the space charge density $\varrho(\boldsymbol{r})$ *or the surface charge density* σ changes the empty space. An electric vector field

$$\boldsymbol{E}(\boldsymbol{r}) = \boldsymbol{F}(\boldsymbol{r})/q$$

is created by the charges. Its amount and direction is in every point defined by the total force $\boldsymbol{F}(\boldsymbol{r})$ on the test charge q.

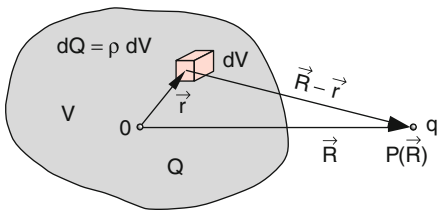


Fig. 1.8 Force on a test charge q by a continuous charge distribution

This field can be illustrated by field lines. Their tangent in the point P points always in the direction of the field strength.

The direction of the field strength is defined by the force on a positive point charge. It points from a positive field charge away to a negative field charge.

Figures 1.9 1.10 and 1.11 show a few examples of electric fields created by point charges.

To illustrate the determination of the field of a surface charge we start with the calculation of the field of an infinite flat sheet of homogeneous charge density σ (Fig. 1.12).

The charge $\mathrm{d}Q = \sigma\,\mathrm{d}A$ exerts a force

$$\mathrm{d}\boldsymbol{F} = \frac{q}{4\pi\varepsilon_0}\frac{\sigma\cdot\mathrm{d}A}{b^2}\hat{\boldsymbol{b}}, \tag{1.7a}$$

on the test charge q at a distance b.

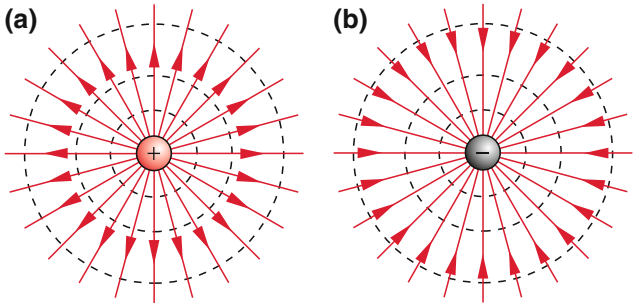


Fig. 1.9 Electric field lines (*red*) produced by a positive and a negative point charge and equipotential lines (*black dashed*)

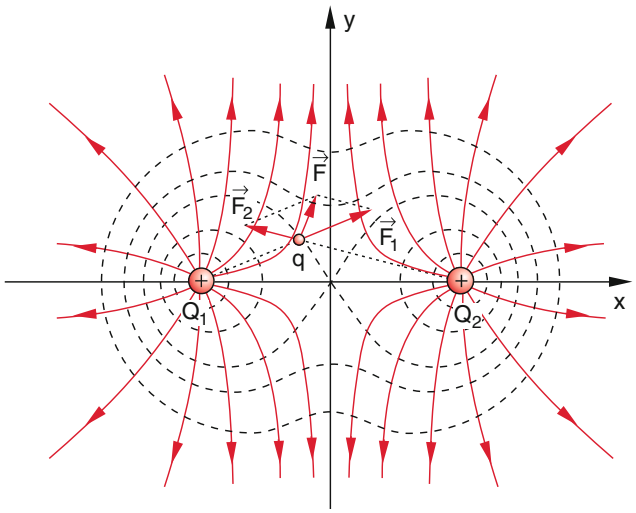


Fig. 1.10 Electric field lines (*red*) and equipotiental lines (*dashed black*) of two positive field charges Q_1 and Q_2

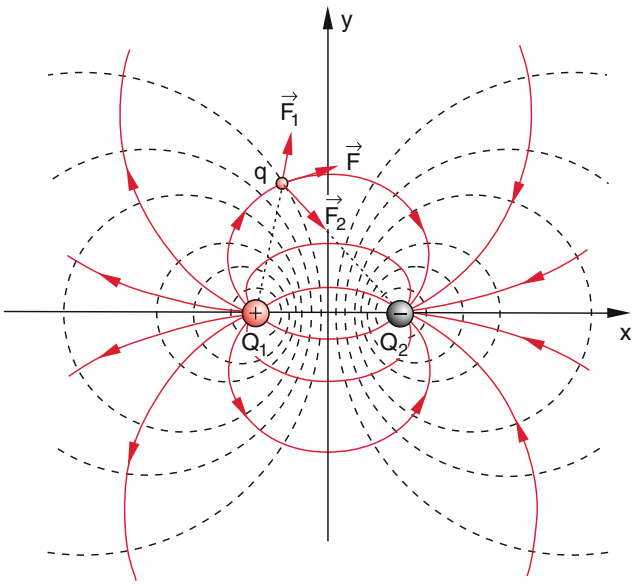


Fig. 1.11 Electric field lines and equipotential lines of an electric dipole

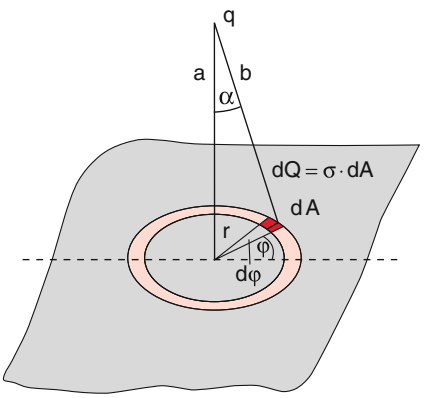


Fig. 1.12 Illustration of the force on a test charge q in the electric field of a surface charge

We split this force into a horizontal component $\mathrm{d}F\cdot\sin\alpha$ parallel to the surface, and a vertical component $\mathrm{d}F\cdot\cos\alpha$ normal to the surface. By integration over the angle φ (see Fig. 1.12) and using the surface element $\mathrm{d}A = r\,\mathrm{d}\varphi\,\mathrm{d}r$ and $b = a/\cos\alpha$ yields the contribution of the vertical component.

$$\begin{aligned}\mathrm{d}F_\mathrm{v} &= \frac{2\pi r\,\mathrm{d}r}{4\pi\varepsilon_0 b^2}\,q\cdot\sigma\cdot\cos\alpha \\ &= \frac{q\cdot\sigma}{2\varepsilon_0 a^2}\left(\cos^3\alpha\right)\cdot r\,\mathrm{d}r,\end{aligned} \tag{1.7b}$$

which is generated by the charges on the circular ring with the area $2\pi r\cdot\mathrm{d}r$ (bright red area in Fig. 1.12).

The amount of the horizontal component becomes zero because of the rotational symmetry i.e. each two opposing components compensate. With $a = \tan\alpha$ and $dr/d\alpha = a/\cos^2\alpha$ we reformulate dF_v *as*

$$dF_v = -\sin\alpha \cdot da$$

Integration over the infinite surface yields the total force onto q that is equivalent to the integration over the angle α from 0 to $\pi/2$.

$$F = \int_0^{\pi/2} dF_v = \frac{q \cdot \sigma}{2\varepsilon_0} \tag{1.7c}$$

The force $\boldsymbol{F}$ is always perpendicular to the plate and therefore also the electric field $\boldsymbol{E} = \boldsymbol{F}/q$ with $\boldsymbol{E} = \sigma/\varepsilon_0$). Furthermore the electric field is independent of the normal distance a from the charged surface. Such a field with a spatially constant vector $\boldsymbol{E}$ is called a *homogeneous field*.

If the charged plate has finite dimensions D then boundary effects disturb the homogeneity of the field. They can be reduced by placing a second plate at a distance d parallel to the existing one. Both plates are charged with $|Q| = |Q_1| = |Q_2|$ but of opposite sign. The distance d between the plates is small compared to their extension ($d \ll D$). Inside of such a charged parallel plate capacitor (see Fig. 1.13 and Sect. 1.5.2) the force on a charge q is therefore

$$\boldsymbol{F} = \frac{\sigma q}{\varepsilon_0}\hat{\boldsymbol{x}}, \quad \hat{\boldsymbol{x}} = \boldsymbol{x}/|\boldsymbol{x}|. \tag{1.8a}$$

The electric field strength $\boldsymbol{E} = \boldsymbol{F}/q$ inside the parallel plate capacitor is then

$$\boldsymbol{E} = \frac{\sigma}{\varepsilon_0}\hat{\boldsymbol{x}}. \tag{1.8b}$$

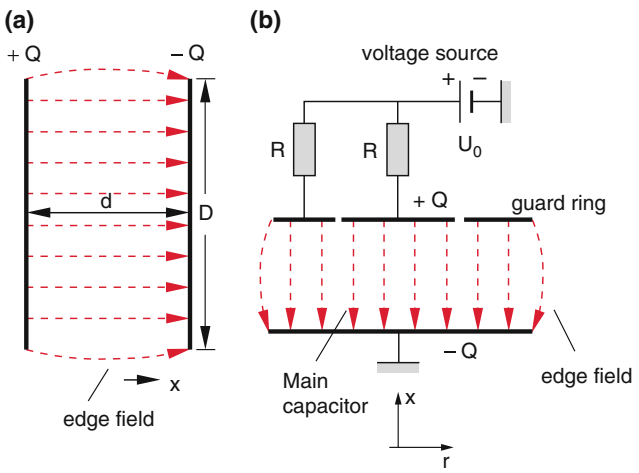


Fig. 1.13 **a**) Electric field of a parallel plate capacitor with edge effects **b**) compensation of edge effects

its amount is

$$E = \frac{\sigma}{\varepsilon_0} \tag{1.8c}$$

Because amount and direction of the vector field are constant within the capacitor, the electric field E is homogeneous.

At the edges of the plates the field is inhomogeneous. That can be minimized by a separate, isolated ring around the plates, with the same potential as the capacitor (guard ring) (Fig. 1.13b).

1.2.2 Electric Flux; Charges as Sources of Electric Fields

We consider a surface that encloses a volume with point charges or with a volume charge density ϱ. The electric field lines of these charges penetrate the surface A. We denote an area element $d\boldsymbol{A}$ of this surface by the outward normal vector $d\boldsymbol{A}$ (Fig. 1.14a). The electric flux $d\Phi_{el}$ through $d\boldsymbol{A}$ is defined by the scalar product

$$d\Phi_{el} = \boldsymbol{E} \cdot d\boldsymbol{A} \tag{1.9a}$$

and is a measure of the number of electric field lines through $d\boldsymbol{A}$. We get the total electric flux through A by integration

$$\Phi_{el} = \int \boldsymbol{E} \cdot d\boldsymbol{A}. \tag{1.9b}$$

A point charge at the center of a sphere with surface A creates the Coulomb field

$$\boldsymbol{E} = Q/(4\pi\varepsilon_0 r^2)\hat{\boldsymbol{r}}$$

and the electric flux through A

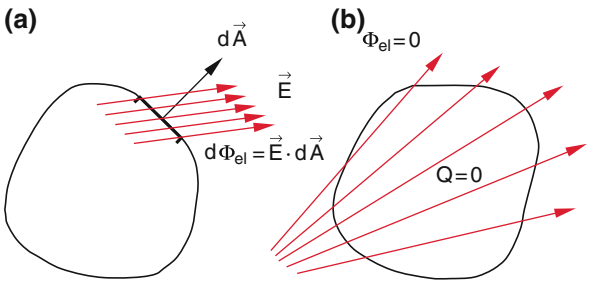


Fig. 1.14 **a**) Illustration of the electric flux through a surface element dA **b**) electric flux through a closed surface

$$\Phi_{\mathrm{el}} = \frac{Q}{4\pi\varepsilon_0}\int \frac{\hat{\boldsymbol{r}}}{r^2}\mathrm{d}\boldsymbol{A} = \frac{Q}{4\pi\varepsilon_0}\int \mathrm{d}\Omega = Q/\varepsilon_0,$$

because $\mathrm{d}A/r^2$ is the solid angle $\mathrm{d}\Omega$ and the integral over the solid angle equals 4π.

With Gauss' theorem it can be shown mathematically that for every closed surface A

$$\Phi_{\mathrm{el}} = \int_A \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{A} = \int_{V(A)} \mathrm{div}\,\boldsymbol{E}\,\mathrm{d}V.$$

is valid [3].

From the result of the special case above the validity for the general case

$$\Phi_{\mathrm{el}} = \frac{1}{\varepsilon_0}Q = \frac{1}{\varepsilon_0}\int \varrho\,\mathrm{d}V \quad (1.10)$$
$$\Rightarrow \mathrm{div}\,\boldsymbol{E} = \varrho/\varepsilon_0$$

can be deduced. In words

> The spatially distributed charges are sources (if $\varrho > 0$) or rather sinks (if $\varrho < 0$) of the electric field.

Note The total electric flux through a closed surface is independent of the form of the surface and the distribution of the charges $\varrho(\boldsymbol{r})$ but depends solely on the total charge Q included by the surface A.

In the model of field lines all field lines originate on a positive charge and terminate on a negative charge (see Fig. 1.11). If the surface A encloses a positive charge Q (or an excess charge $\Delta Q > 0$) then $\Phi_{\mathrm{el}} > 0$ i.e. more field lines come out of the enclosed volume than enter the volume. If the total charge inside the volume is zero then also $\Phi_{\mathrm{el}} = 0$ and the number of field lines entering the surface is the same as the number of lines leaving the surface (Figs. 1.14b and 1.15).

1.3 The Electrostatic Potential

In order to bring a charge q in an electric field $\boldsymbol{E}$ from the position P_1 to P_2 (Fig. 1.16) work has to be done (see Vol. 1, (2.35))

$$W = \int_{P_1}^{P_2} \boldsymbol{F}\cdot\mathrm{d}\boldsymbol{s} = q\cdot\int_{P_1}^{P_2} \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s}. \quad (1.11)$$

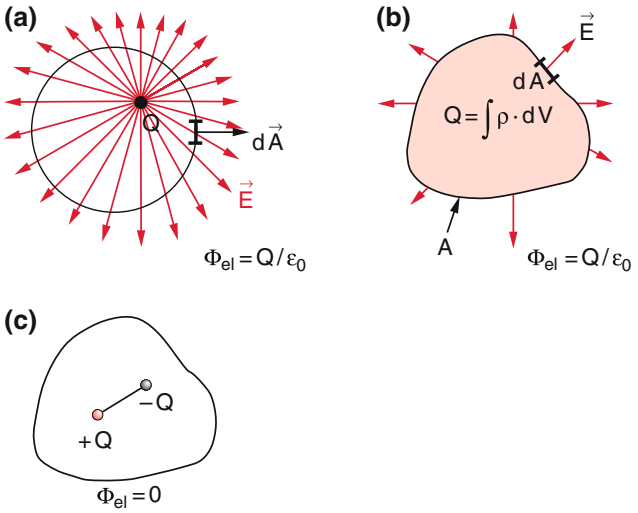


Fig. 1.15 **a**) Electric flux through a closed surface including a positive point charge Q **b**) Electric flux produced by a volume charge **c**) The electric flux through a closed surface including an electric dipole is zero

Example

In the field of a point charge Q a test charge q is brought from the distance r_1 to the distance r_2. The necessary work is

$$W = \frac{qQ}{4\pi\varepsilon_0}\int_{r_1}^{r_2}\frac{\mathrm{d}r}{r^2} = \frac{qQ}{4\pi\varepsilon_0}\left(\frac{1}{r_1}-\frac{1}{r_2}\right)$$

If the distance between charges of equal sign is increased $(r_2 > r_1)$ then $W > 0$ i.e. we gain energy at the expenses of potential energy. When decreasing the distance between the repelling charges we have to supply energy $(W < 0)$.

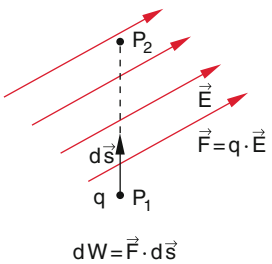


Fig. 1.16 Illustration of the work necessary to bring a test charge q in an electric field from P_1 to P_2

1.3.1 Potential and Voltage

When we discussed the gravity potential (Vol. 1, Sect. 2.7) it has been shown that in a conservative force field the work-integral is independent of the path and depends solely on the end points P_1 and P_2. Because the electric field is conservative as the gravity field we can attribute to each point of space an unambiguously defined function

$$\phi(P) = \int_P^\infty \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \tag{1.12}$$

which is called the electric potential at the point P. It is usually set to zero at infinity: $\phi(r = \infty) = 0$.

The product $q \cdot \phi(r)$ gives the work that has to be done resp. which is gained if the charge q is transferred from the point P to infinity.

The potential difference between two points P_1 and P_2

$$U = \phi(P_1) - \phi(P_2) = \int_{P_1}^{P_2} \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \tag{1.13}$$

is called the **electric voltage** U (Fig. 1.17).

A charge q that is transferred in space and overcomes the electric potential difference U suffers a change of its potential energy

$$\Delta E_{\mathrm{pot}} = -qU \tag{1.14}$$

Because the total energy $E = E_{\mathrm{kin}} + E_{\mathrm{pot}}$ is constant the kinetic energy must change by

$$\Delta E_{\mathrm{kin}} = -\Delta E_{\mathrm{pot}} = qU. \tag{1.14a}$$

The unit of the potential difference (voltage) is 1 V (V).

$$[U] = \left[E_{\mathrm{pot}}/q\right] = 1\,\mathrm{N\,m/(A\,s)} = 1\,\mathrm{V\,A\,s/(A\,s)} = 1\,\mathrm{V}.$$

In atomic physics it is more convenient to use the smaller unit of energy 1 electron volt (eV). This is the energy gained by an electron if it is accelerated by a potential difference $U = \Delta\phi = 1\,\mathrm{V}$. According to (1.14a) is

$$1\,\mathrm{eV} = 1.602 \times 10^{-19}\,\mathrm{C} \cdot 1\,\mathrm{V} = 1.602 \times 10^{-19}\,\mathrm{J}.$$

Examples

1. A hot cathode in an evacuated tube emits electrons with an initial speed v_0. The voltage U between cathode and anode (Fig. 1.18) accelerates the electrons. Their energy at the anode is then

$$E_{\mathrm{kin}} = \frac{m}{2}v^2 = \frac{m}{2}v_0^2 + e \cdot U.$$

They hit the anode with a velocity $v = \sqrt{v_0^2 + 2eU/m}$. Because generally $v_0 \ll v$ we can approximate $v \approx \sqrt{2eU/m}$. With $U = 50\,\mathrm{V}$ is $v = 4 \times 10^6\,\mathrm{m/s}$, which is about 1.3% of the speed of light.

2. How much energy has to be spent to ionize a hydrogen atom, i.e. to bring the electron from the distance r_1 from the proton to infinity.

$$W = \frac{-e^2}{4\pi\varepsilon_0}\int_{r_1}^{\infty}\frac{\mathrm{d}r}{r^2} = \frac{-e^2}{4\pi\varepsilon_0 r_1}.$$

With the numerical values $e = 1.6 \times 10^{-19}\,\mathrm{C}$, $\varepsilon_0 = 8.85 \times 10^{-12}\,\mathrm{C/Vm}$ and $r_1 = 5 \times 10^{-11}\,\mathrm{m}$ we get $W = -27\,\mathrm{eV}$. The experimental value is $W_{\mathrm{exp}} = -13.5\,\mathrm{eV}$. The discrepancy comes from neglecting the kinetic energy in the ground state of the hydrogen atom. The mean value of the kinetic energy in a force field $F \propto 1/r^2$ is

$$\langle E_{\mathrm{kin}} \rangle = -\langle 2E_{\mathrm{pot}} \rangle \quad \text{(Virial theorem)}.$$

This can be readily verified for a circular motion of a charge q on a circle with radius r by equating centripetal and Coulomb force.

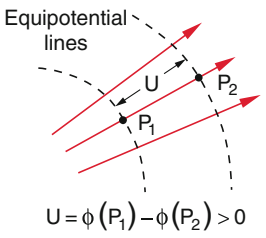


Fig. 1.17 Equipotential lines and electric field lines

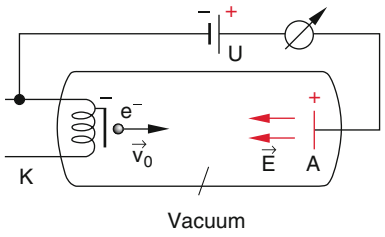


Fig. 1.18 Electrons emitted by a hot cathode are accelerated by the voltage U

1.3.2 Potential Equation

It follows from the definition of the electric potential

$$\phi(P) = \int_P^\infty \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s}$$

exactly as for the gravity potential that the field strength $\boldsymbol{E}$ can be written as the gradient

$$\boldsymbol{E} = -\mathbf{grad}\,\phi(x,y,z) = -\nabla\phi. \tag{1.15}$$

The electrostatic field can either be described by a scalar potential function $\phi(x,y,z)$ or by the vector field $\boldsymbol{E}(x,y,z)$. The scalar potential function assigns to each point in space a scalar quantity while the vector field assigns to each point in space the triple $\{E_x, E_y, E_z\}$ that defines magnitude and direction of the electric field in this point.

From (1.10) it follows

$$\mathrm{div}\,\boldsymbol{E} = -\mathrm{div}\,\mathbf{grad}\,\phi = -\Delta\phi = \varrho/\varepsilon_0, \tag{1.16}$$

where Δ represents the Laplace operator (see Vol. 1, Chap. 13).

The equation

$$\Delta\phi = -\varrho/\varepsilon_0 \tag{1.16a}$$

is called *Poisson equation*.

The integration of this differential equation makes it possible to determine the potential $\phi(x,y,z)$ and the electric field $\boldsymbol{E}(x,y,z)$ if the charge distribution $\varrho(x,y,z)$ is given.

The constants of integration are determined by suitable boundary conditions. In a space without charges the Poisson Eq. (1.16a) simplifies to the Laplace equation

$$\mathrm{div}\,\mathbf{grad}\,\phi = \Delta\phi = 0 \tag{1.16b}$$

Equation (1.16) plays an important role in electrostatics comparable to Newton's equation of motion $\boldsymbol{F} = m\boldsymbol{a}$ in mechanics.

If the distribution of charges $\varrho(\boldsymbol{r})$ is known the potential $\phi(\boldsymbol{r})$ and the field strength $\boldsymbol{E}(\boldsymbol{r})$ always can be determined, at least numerically.

In Sect. 1.3.4 we will illustrate the calculation of potential and electric field for some examples.

1.3.3 Equipotential Surfaces

Surfaces with constant potential $\phi(\boldsymbol{r})$ are called *equipotential surfaces*. Analysis taught us that the gradient (in this case the electric field $\boldsymbol{E} = -\mathrm{grad}\,\phi$) is perpendicular to an equipotential surface at any point of the surface. One can imagine the equipotential lines similar to contour lines of maps. Contour lines connect all points with a given altitude (= distance to the mean sea level).

The mathematical expression for a three-dimensional contour map is a scalar valued function $z(x, y)$ describing the distance z to the x-y-plane ($z = 0$). If we describe the surface of mountains by the set of all points (x,y,z) with height $z = h(x,y)$ then a contour line is the subset $h(x,y) = \text{constant}$.

At any point of a contour line the gradient points in the direction of steepest ascent and is always normal to the contour line.

The contour lines also can be seen as lines of constant potential energy E_p. At any point the field lines are parallel to the forces and therefore are normal to the contour lines. For electric field lines the behavior is completely analogous.

Therefore the equipotential surfaces are surfaces perpendicular to the field lines (Figs. 1.9, 1.10 and 1.11).

Moving a charge on an equipotential surface requires no work.

$$W = q \cdot \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} \equiv 0 \quad \text{because } \boldsymbol{E} \perp \mathrm{d}\boldsymbol{s}.$$

Examples

1. The equipotential surface of the Coulomb field of a positive charge is a sphere about the charge at its center (Fig. 1.9). If there are more than one point charges things become more complicated. In the case of two point charges we get the equipotential surface of Figs. 1.10 and 1.11.
2. In the homogeneous field of a parallel plate capacitors (Fig. 1.13) the equipotential surfaces are planes parallel to the plates.
3. In electrostatics with its charges at rest all surfaces of conductors are equipotential surfaces. All field lines are perpendicular to the surface of a conductor. This is no longer valid if there are electric currents through the conductors (see Sect. 2.2.2).

1.3.4 Special Distributions of Charges

1.3.4.1 Charged Hollow Spheres

The homogeneously charged surface of a conducting hollow sphere of radius R has the surface charge density σ and the total charge $Q = 4\pi R^2\sigma$. According to (1.9b) the electric flux through the surface of a concentric sphere of radius $r > R$ is

$$\Phi_{el} = \int \boldsymbol{E} \cdot d\boldsymbol{A} = E \cdot 4\pi r^2 = Q/\varepsilon_0 \Rightarrow \boldsymbol{E} = \frac{Q}{4\pi\varepsilon_0 r^2}\hat{\boldsymbol{r}},$$

because of symmetry arguments $\boldsymbol{E}$ points radially outward, i.e. $\boldsymbol{E} \| d\boldsymbol{A} \| \hat{\boldsymbol{r}}$.

The charged surface of a sphere of radius R acts at $r > R$ like a point charge Q in its center.

We obtain the potential at a distance $r > R$ to the center of the hollow sphere from

$$\phi(r) = \int_r^\infty E \cdot dr = \frac{Q}{4\pi\varepsilon_0 r} \Rightarrow |\boldsymbol{E}(\boldsymbol{r})| = \frac{\phi(r)}{r}.$$

Since the surface of a conductor is an equipotential surface the electric field strength rises with decreasing radius of curvature if the potential ϕ has a fixed amount.

An arbitrary closed surface completely inside the sphere ($r < R$) does not surround any charge. Because for each of these surfaces is

$$\Phi_{el} = \int \boldsymbol{E} \cdot d\boldsymbol{A} = 0,$$

it follows that $\boldsymbol{E} = 0$ inside the sphere.

Inside the homogeneously charged hollow sphere is no electric field. The potential inside the sphere is constant (Fig. 1.19).

1.3.4.2 Charged Solid Sphere

Electric field and potential of a homogeneously charged not conducting solid sphere with charge $Q = \frac{4}{3}\pi R^3\varrho$ are given analogous to Vol. 1, Sect. 2.9 (see Fig. 1.20 and Problem (1.8)):

For $r \geq R$ we get

$$\boldsymbol{E} = \frac{Q}{4\pi\varepsilon_0 r^2}\hat{\boldsymbol{r}}; \quad \phi = \frac{Q}{4\pi\varepsilon_0 r}, \tag{1.17a}$$

resp. for $r \leq R$

$$\boldsymbol{E} = \frac{Qr}{4\pi\varepsilon_0 R^3}\hat{\boldsymbol{r}}; \quad \phi = \frac{Q}{4\pi\varepsilon_0 R}\left(\frac{3}{2} - \frac{r^2}{2R^2}\right). \tag{1.17b}$$

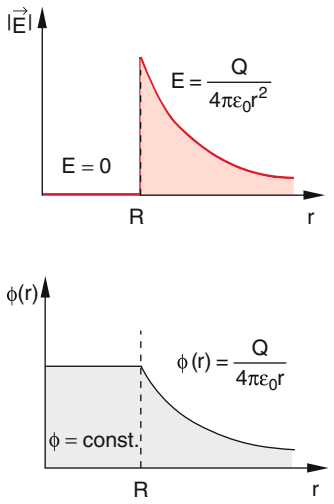


Fig. 1.19 Electric field and potential inside and outside of a charged hollow sphere

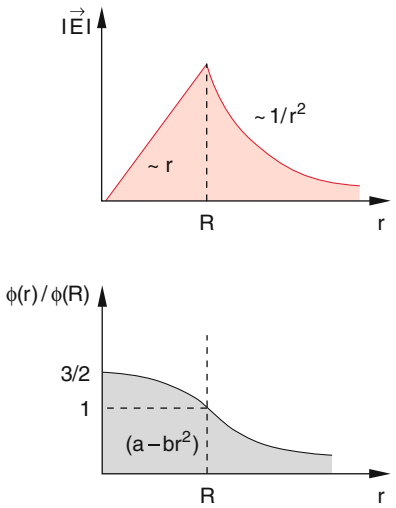


Fig. 1.20 Electric field and normalized potential of a uniformly charged non-conducting solid sphere

1.3.4.3 Charged Rod

As a further example we will calculate field and potential of a charged infinitely long rod of radius R (Fig. 1.21). The charge per unit length is $\lambda = Q/L = \pi R^2 \cdot \varrho$ Again for symmetry arguments the field strength $\boldsymbol{E}$ at a point P with a distance r from the rod axis is directed radially outward. The electric flux through the surface of a coaxial cylinder of radius r and length L is for $r \geq R$

$$\begin{aligned} \Phi_{\text{el}} &= \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} = E \cdot 2\pi r \cdot L \\ &= \frac{Q}{\varepsilon_0} = \frac{\lambda}{\varepsilon_0} \cdot L \Rightarrow |\boldsymbol{E}| = \frac{\phi_{\text{el}}}{2\pi r \cdot L} \\ &\Rightarrow \boldsymbol{E} = \frac{\lambda}{2\pi\varepsilon_0 r}\hat{\boldsymbol{r}}. \end{aligned} \tag{1.18a}$$

(λ = Q/L= charge per unit length).

For $r \leq R$ we have $Q = \lambda \cdot L \cdot \pi r^2/(\pi R^2)$

$$\begin{aligned} \Rightarrow \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} &= E \cdot 2\pi r \cdot L = \frac{\lambda \cdot L \cdot r^2}{\varepsilon_0 R^2} \\ \Rightarrow \boldsymbol{E} &= \frac{\lambda \boldsymbol{r}}{2\varepsilon_0 \pi R^2}. \end{aligned} \tag{1.18b}$$

With the boundary condition $\phi(R) = 0$ we get for $r \geq R$ the electric potential

$$\phi(r) = -\frac{\lambda}{2\pi\varepsilon_0} \ln \frac{r}{R} \tag{1.18c}$$

and for $r \leq R$

$$\phi(r) = \frac{\lambda}{4\pi\varepsilon_0}\left(1 - \frac{r^2}{R^2}\right). \tag{1.18d}$$

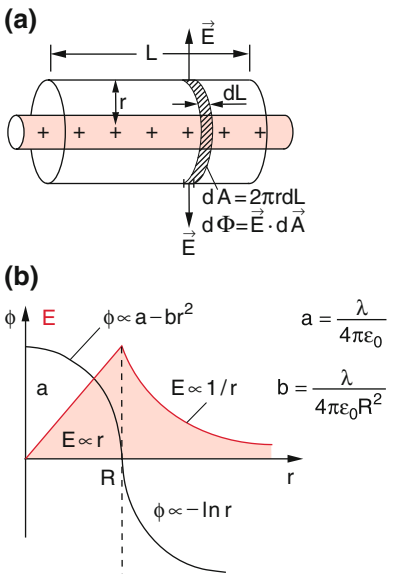


Fig. 1.21 Electric field and potential of an infinitely long charged rod

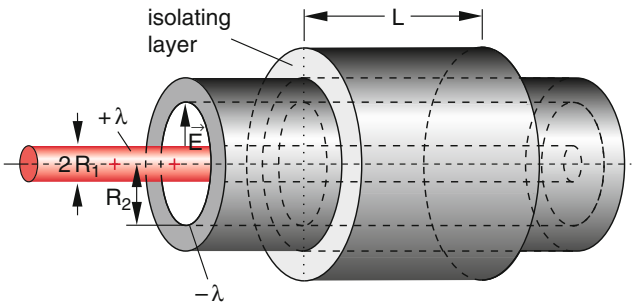


Fig. 1.22 Coaxial cable

Question: Why is here the boundary condition $\phi(\infty) = 0$ not useful?

1.3.4.4 Coaxial Cable

A coaxial cable is an arrangement of a conducting wire of radius R_1 surrounded by a coaxial hollow conducting cylinder of radius R_2 (Fig. 1.22). Both conductors carry the same amount of charge density but of opposite sign, $\lambda_1 = -\lambda_2$.

For $r > R_2$ we have

$$\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} = 0 \Rightarrow E = 0,$$

because the total charge insides a cylinder of radius $r > R_2$ is zero.

Let be $R_1 \leq r \leq R_2$.

The field caused by the outer cylinder is zero, because at $r < R_2$ there are no (negative) charges. The field caused by the inner wire is

$$\boldsymbol{E} = \frac{\lambda}{2\pi\varepsilon_0 r}\hat{\boldsymbol{r}}.$$

as we have derived in the previous example.

1.4 Electric Multipoles

The Poisson equation (1.16) is a linear equation in E and ϕ, i.e. the Poisson equation depends only linearly on the electric field $E(\boldsymbol{r})$ as well as on its components and the potential $\phi(\boldsymbol{r})$. It follows that the Coulomb potentials $\phi(\boldsymbol{r})$ generated by the charges Q_i distributed in space can be linearly superimposed at the point P.

Therefore N point charges $Q_i(\boldsymbol{r}_i)$ (Fig. 1.23) produce at the point P the total potential

$$\phi(\boldsymbol{R}) = \frac{1}{4\pi\varepsilon_0} \sum_{i=1}^{N} \frac{Q_i}{|\boldsymbol{R} - \boldsymbol{r}_i|}, \tag{1.19}$$

where $\boldsymbol{R}$ is the position vector of P and $\boldsymbol{r}_i$ the position vectors of the charges Q_i.

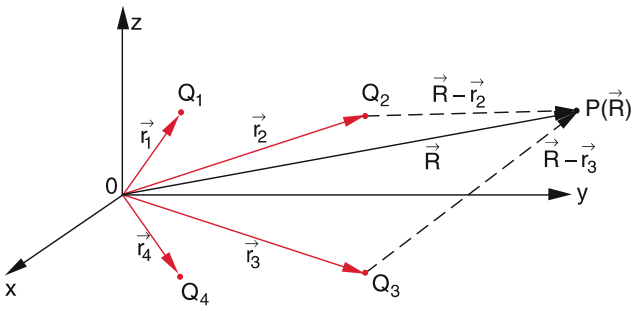


Fig. 1.23 Potential generated at the observation point $P(R)$ by four point charges Q_i

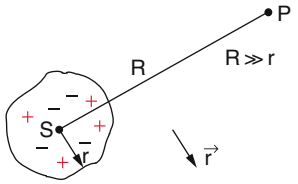


Fig. 1.24 Potential of a charge distribution in a volume $V(r)$ at a far distant point $P(R)$ with $R >> r$

Now we consider a continuous distribution of space charges with the density $\varrho(\boldsymbol{r})$ (Fig. 1.24). Because of $Q = \int \varrho \, \mathrm{d}V$ we have

$$\phi(\boldsymbol{R}) = \frac{1}{4\pi\varepsilon_0} \int_V \frac{\varrho(\boldsymbol{r}) \, \mathrm{d}V}{|\boldsymbol{R} - \boldsymbol{r}|}, \tag{1.20}$$

where the origin of the coordinate system is often chosen at the center of charges S.

The integral (1.20) over an arbitrary charge distribution often has no analytic solution but if the distances R between the center of charges R = 0 and the point $P(\boldsymbol{R})$ is large compared to the extension of the space charges one can expand the potential $\phi(\boldsymbol{R})$ in a Taylor series and integrate each term separately. The origin of the coordinate system is either the center of the charge distribution S of one sign or the midpoint between the centers of positive and negative charges. Then $\phi(R)$ is expanded in terms of $r/R \ll 1$.

This so called *multipole expansion* divides the potential of the charge distribution into sums $\phi_n(R)$ that are generated by point charges (monopoles), by pairs of point charges (dipoles), by pairs of dipoles (quadrupoles) and so on. Each of these contributions decreases with another power R^{-n} depending on the distance R between field point and center of charges S (Fig. 1.25). This model has been very useful for example to calculate the interaction of atoms and molecules. We get better insight into the distribution of charges.

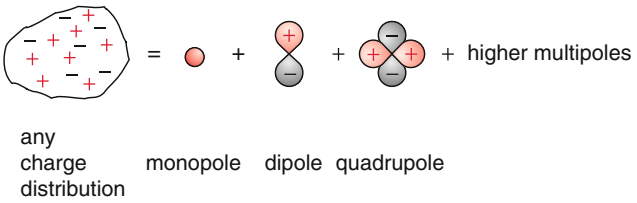


Fig. 1.25 Illustration of multipole expansion

Now we will treat the spatial distributions of potentials and of fields for some simple multipoles to illustrate the general multipole expansion discussed in Sect. 1.4.3 by these concrete examples.

1.4.1 The Electric Dipole

An electric dipole consists of two charges of opposite signs but equal amounts $Q_1 = Q = -Q_2$ at a distance d (Fig. 1.26).

It is characterized by its dipole moment

$$\boldsymbol{p} = Q \cdot \boldsymbol{d},$$

Its direction is defined by pointing from the negative charge to the positive charge where $\boldsymbol{d}$ is the distance between $-Q$ to $+Q$.

Field strength $\boldsymbol{E}(\boldsymbol{R})$ and potential $\phi(\boldsymbol{R})$ at any point $P(\boldsymbol{R})$ is given by superposition of the fields of both point charges. We chose the origin at the midpoint S between $+Q$ and $-Q$. The calculation of the field is easier to accomplish if we first calculate the potential and then apply the gradient operator to the potential. Let $\boldsymbol{r}_1 = \boldsymbol{R} - \boldsymbol{d}/2$ and $\boldsymbol{r}_2 = \boldsymbol{R} + \boldsymbol{d}/2$ then we get

$$\phi_{\mathrm{D}}(\boldsymbol{R}) = \frac{1}{4\pi\varepsilon_0} \left(\frac{Q}{|\boldsymbol{R} - \boldsymbol{d}/2|} - \frac{Q}{|\boldsymbol{R} + \boldsymbol{d}/2|} \right). \tag{1.21}$$

At sufficiently large distance from the dipole ($R \gg d$) we can cut the Taylor series expansion after the linear term

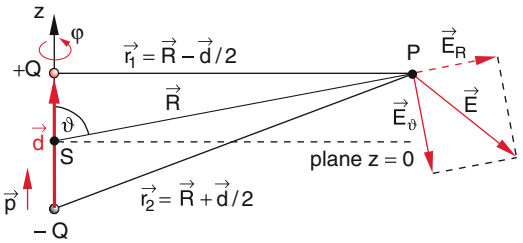


Fig. 1.26 Electric field of an electric dipole

$$\frac{1}{|\boldsymbol{R} \pm \boldsymbol{d}/2|} = \frac{1}{R} \cdot \frac{1}{\sqrt{1 \pm \frac{\boldsymbol{R}\cdot\boldsymbol{d}}{R^2} + \frac{d^2}{4R^2}}} = \frac{1}{R}\left(1 \mp \frac{1}{2}\frac{\boldsymbol{R}\cdot\boldsymbol{d}}{R^2} + \cdots\right) \tag{1.22}$$

and get an approximation to the potential of a dipole at far distances.

$$\phi_{\mathrm{D}}(\boldsymbol{R}) = \frac{Q}{4\pi\varepsilon_0} \cdot \frac{\boldsymbol{d}\cdot\boldsymbol{R}}{R^3} = \frac{\boldsymbol{p}\cdot\boldsymbol{R}}{4\pi\varepsilon_0 R^3} = \frac{p\cdot\cos\vartheta}{4\pi\varepsilon_0 R^2}. \tag{1.23}$$

Because of $\mathbf{grad}\,(1/r) = -\boldsymbol{r}/r^3$ we can rewrite the dipole potential as the product of distance d between the charges and the gradient of the monopole potential (Coulomb potential)

$$\phi_{\mathrm{D}}(\boldsymbol{R}) = -\frac{Q}{4\pi\varepsilon_0}\boldsymbol{d}\cdot\nabla\left(\frac{1}{R}\right) = -\boldsymbol{d}\cdot\mathbf{grad}\,\phi_{\mathrm{M}}(\boldsymbol{R}) \tag{1.24}$$

This illustrates that with increasing distance R the potential $\phi_{\mathrm{D}}(R) \propto 1/R^2$ decreases faster than the potential $\phi_{\mathrm{M}}(R) \propto 1/R$ of a monopole. The reason for it is that the opposing potentials of $+Q$ and $-Q$ more and more compensate with increasing distance. In the symmetry plane $z = 0$ is $\vartheta = 90°$ and therefore everywhere $\phi_{\mathrm{D}} \equiv 0$.

The electric field $\boldsymbol{E} = -\mathbf{grad}\,\phi_{\mathrm{D}}$ can be calculated from (1.23) using

$$\mathbf{grad}\,\phi_{\mathrm{D}} = \frac{Q}{4\pi\varepsilon_0}\left\{(\boldsymbol{d}\cdot\boldsymbol{R})\mathbf{grad}\frac{1}{R^3} + \frac{1}{R^3}\mathbf{grad}(\boldsymbol{d}\cdot\boldsymbol{R})\right\}$$

Because of $\mathbf{grad}\,1/R^3 = -3\boldsymbol{R}/R^5$ and $Q\boldsymbol{d}\cdot\boldsymbol{R} = p\cdot R\cdot\cos\vartheta$ and $Q\,\mathbf{grad}(\boldsymbol{d}\cdot\boldsymbol{R}) = \boldsymbol{p}$ we get

$$\boldsymbol{E}(\boldsymbol{R}) = \frac{1}{4\pi\varepsilon_0 R^3}(3p\widehat{\boldsymbol{R}}\cdot\cos\vartheta - \boldsymbol{p}). \tag{1.25a}$$

The field can be best illustrated in polar coordinates (R, ϑ, φ) because of the cylindrical symmetry of the problem.

From

$$\boldsymbol{E} = -\mathbf{grad}\,\phi = -\left\{\frac{\partial\phi}{\partial R}, \frac{1}{R}\frac{\partial\phi}{\partial\vartheta}, \frac{1}{R\sin\vartheta}\frac{\partial\phi}{\partial\varphi}\right\}, 0$$

we get with (1.23)

$$E_R = \frac{2p\cdot\cos\vartheta}{4\pi\varepsilon_0 R^3}, \quad E_\vartheta = \frac{p\cdot\sin\vartheta}{4\pi\varepsilon_0 R^3}, \quad E_\varphi = 0. \tag{1.25b}$$

The field does not depend on the azimuth angle φ and is therefore cylindrical symmetric about the dipole axis.

Figure 1.11 shows the electric field in the x-y-plane that contains the axis of the dipole. Figure 1.26 shows the components E_R and E_ϑ of the field strength.

Electric field E and potential ϕ of the dipole have cylindrical symmetry about the dipole axis which we choose as the z-axis.

1.4.1.1 The Dipole in a Homogeneous Electric Field

In an external electric field $E(r)$ the electric dipole has the potential energy (Fig. 1.27)

$$W_{\mathrm{pot}} = Q\phi_1 - Q\phi_2 = Q(\phi_1 - \phi_2), \tag{1.26}$$

which becomes zero if both charges $+Q$ and $-Q$ are situated on an equipotential surface i.e. the dipole axis is normal to $\boldsymbol{E}$.

For an arbitrary position of the dipole the homogeneous electric field acts with the forces $\boldsymbol{F}_1 = Q\cdot\boldsymbol{E}$ and $\boldsymbol{F}_2 = -Q\cdot\boldsymbol{E}$ on the charges Q and $-Q$. Because of $\boldsymbol{r}_1 - \boldsymbol{r}_2 = \boldsymbol{d}$ the forces generated a torque

$$\boldsymbol{D} = Q(\boldsymbol{r}_1\times\boldsymbol{E}) - Q(\boldsymbol{r}_2\times\boldsymbol{E}) = (Q\cdot\boldsymbol{d})\times\boldsymbol{E} = \boldsymbol{p}\times\boldsymbol{E}$$

that is normal to $\boldsymbol{d}$ and $\boldsymbol{E}$ and therefore can be written in vector mode

$$\boldsymbol{D} = \boldsymbol{p}\times\boldsymbol{E}. \tag{1.27}$$

The potential energy of the dipole in an external homogeneous field results from (1.26) because of $\phi_1 - \phi_2 = \mathbf{grad}(\phi\cdot\boldsymbol{d})$ and $\boldsymbol{E} = -\mathbf{grad}\,\phi$ as

$$W_{\mathrm{pot}} = -\boldsymbol{p}\cdot\boldsymbol{E} \tag{1.28}$$

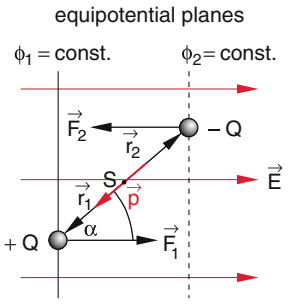


Fig. 1.27 Electric dipole in a homogeneous electric field

The potential energy has a minimum if p and $\boldsymbol{E}$ are parallel.

The dipole adjusts itself in this position if it is not hindered by other forces.

1.4.1.2 The Dipole in an Inhomogeneous Electric Field

In an inhomogeneous field E the resulting force

$$\begin{aligned} \boldsymbol{F} &= Q \cdot [\boldsymbol{E}(\boldsymbol{r}+\boldsymbol{d}) - \boldsymbol{E}(\boldsymbol{r})] \\ &= Q \cdot \boldsymbol{d} \cdot \frac{\mathrm{d}\boldsymbol{E}}{\mathrm{d}\boldsymbol{r}} = \boldsymbol{p} \cdot \nabla \boldsymbol{E}. \end{aligned} \tag{1.29}$$

acts on the dipole.

The vector gradient of $\boldsymbol{E}$ is a tensor and its scalar product with the vector $\boldsymbol{p}$ yields the vector $\boldsymbol{F}$.

$$\begin{aligned} F_x &= \boldsymbol{p} \cdot \mathbf{grad}\, E_x \\ &= p_x \frac{\partial E_x}{\partial x} + p_y \frac{\partial E_x}{\partial y} + p_z \frac{\partial E_x}{\partial z}, \\ F_y &= \boldsymbol{p} \cdot \mathbf{grad}\, E_y \\ &= p_x \frac{\partial E_y}{\partial x} + p_y \frac{\partial E_y}{\partial y} + p_z \frac{\partial E_y}{\partial z}, \\ F_z &= \boldsymbol{p} \cdot \mathbf{grad}\, E_z \\ &= p_x \frac{\partial E_z}{\partial x} + p_y \frac{\partial E_z}{\partial y} + p_z \frac{\partial E_z}{\partial z}. \end{aligned} \tag{1.29a}$$

The resulting force on a dipole in a homogeneous field is zero. At an arbitrary orientation of $\boldsymbol{p}$ the torque $\boldsymbol{D} = \boldsymbol{p} \times \boldsymbol{E}$ *acts* and turns the dipole into the direction of the field i.e. to its energy minimum. In an inhomogeneous field a force $\boldsymbol{F} = \boldsymbol{p} \cdot \nabla \boldsymbol{E}$ acts on the dipole that turns it into the direction of the field and pulls it into the direction of increasing field strength (Fig. 1.28).

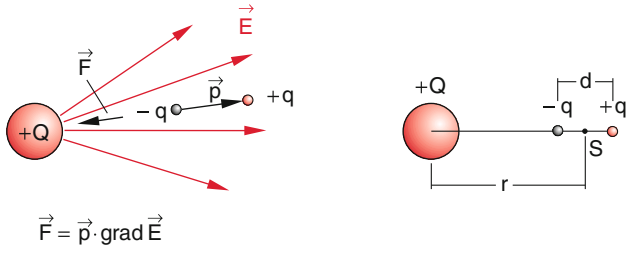


Fig. 1.28 Electric dipole in an inhomogeneous electric field

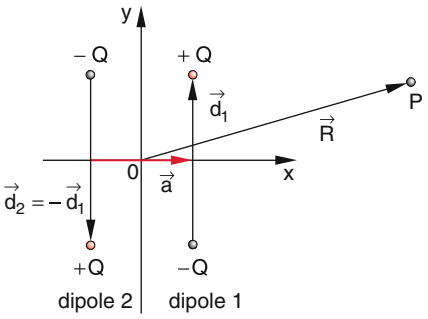


Fig. 1.29 Electric quadrupole

1.4.2 The Electric Quadrupole

Now we arrange two positive and two negative charges in such a way that we have two neighboring antiparallel dipoles at a distance a (Fig. 1.29). So at a large distance R where $R >> a$ and $R >> d$ (distance between the charges) the dipole fields practically cancel each other.

Such an arrangement of four monopoles where the total charge is zero is called a quadrupole The potential is the result of the superposition of two dipole potentials

$$\begin{aligned} \phi_{\mathrm{Q}}(\boldsymbol{R}) &= \phi_{\mathrm{D}}(\boldsymbol{R}+\boldsymbol{a}/2) - \phi_{\mathrm{D}}(\boldsymbol{R}-\boldsymbol{a}/2) \\ &= \boldsymbol{a} \cdot \mathbf{grad}\, \phi_{\mathrm{D}}. \end{aligned} \tag{1.30}$$

From (1.23) we get

$$\phi_{\mathrm{Q}}(\boldsymbol{R}) = \frac{Q}{4\pi\varepsilon_0} \boldsymbol{a} \cdot \mathbf{grad} \left(\frac{\boldsymbol{d} \cdot \boldsymbol{R}}{R^3} \right). \tag{1.31}$$

This shows that the quadrupole potential can be written as a scalar product of distance vector a between the dipoles and gradient of the dipole potential ϕ_{D}. This is analogous to the dipole potential that is equal to the negative scalar product of the distance of the charges and gradient of the monopole potential (1.24). The sign of ϕ_{Q} Q results from the definition of the direction of a.

1.4.3 Multipole Expansion

At a point $P(\boldsymbol{R})$ the potential of an arbitrary distribution of point charges (1.19), surface charges or volume charges (1.20) can be evaluated by a series expansion in powers of r/R (Fig. 1.23). The conditions to do this are: the distance R between the center of the charge distribution and the point $P(R)$ is large compared to r, $r/R \ll 1$ and a sufficient number of terms has to be taken into account in order to acquire the desired accuracy.

The expression

$$\frac{1}{|\boldsymbol{R}-\boldsymbol{r}|}=\frac{1}{\sqrt{(\boldsymbol{R}-\boldsymbol{r})^2}} =\frac{1}{R}\frac{1}{\sqrt{1-(2\boldsymbol{R}\cdot\boldsymbol{r}/R^2)+r^2/R^2}} \tag{1.32}$$

in the sum of (1.19) resp. in the integral of (1.20) can be expanded in a Taylor series (see [3]).

The expansion of the function

$$\begin{aligned} f(x) &= \frac{1}{\sqrt{1-x}} \\ &= f(0)+x\cdot f'(0)+\frac{x^2}{2}f''(0)+\cdots \\ &= 1+\frac{1}{2}x+\frac{3}{8}x^2+\cdots \end{aligned} \tag{1.33}$$

with $x = 2(\boldsymbol{R}\cdot\boldsymbol{r})/R^2 - r^2/R^2$ from (1.32) yields

$$\begin{aligned} &\Rightarrow \sqrt{1-x} = R^{-1}\cdot\sqrt{R^2-2(\boldsymbol{R}\cdot\boldsymbol{r})+r^2}=\frac{1}{2}|\boldsymbol{R}-\boldsymbol{r}| \\ &\frac{1}{|\boldsymbol{R}-\boldsymbol{r}|}=\frac{1}{R}-\boldsymbol{r}\cdot\nabla\frac{1}{R} \\ &\quad+\frac{1}{2}(\boldsymbol{r}\cdot\nabla)(\boldsymbol{r}\cdot\nabla)\frac{1}{R}+\cdots, \end{aligned} \tag{1.34}$$

as you can proof by explicitly carrying out the differentiation of (1.32).

The ∇ operator (also called nabla operator) in (1.34) acts only on R. Inserting (1.34) in (1.19) yields the multipole expansion

$$\begin{aligned} \phi(\boldsymbol{R}) = \frac{1}{4\pi\varepsilon_0}\Bigg[&\frac{1}{R}\sum_{i=1}^{N}Q_i+\frac{1}{R^3}\sum_{i=1}^{N}(Q_i\boldsymbol{r}_i)\boldsymbol{R} \\ &+\frac{1}{R^5}\sum_{i=1}^{N}\frac{Q_i}{2}\big[(3x_i^2-r_i^2)X^2 \\ &+(3y_i^2-r_i^2)Y^2+(3z_i^2-r_i^2)Z^2 \\ &+2(3x_iy_iXY+3x_iz_iXZ \\ &+3y_iz_iYZ)\big]+\ldots\Bigg]. \end{aligned} \tag{1.35}$$

The first term in (1.35) (monopole term) gives the Coulomb potential generated by the total charge at the origin. Therefore this term is zero for neutral charge distributions ($\sum Q_i = 0$) e.g. a neutral atom or molecule. The second term in (1.35) can be written using the electric dipole moment $\boldsymbol{p_i} = Q_i\,\boldsymbol{r_i}$ of the ith charge as $1/R^3\cdot\sum\boldsymbol{p_i}\cdot\boldsymbol{R}$. This dipole term depends not only on the sum of the dipole moments but also on their orientation with respect to the direction $\boldsymbol{R}$ to the observation point P. For a neutral molecule with permanent electric dipole moment, for example NaCl = Na^+ Cl^-, the dipole moment is the leading term in the multipole expansion.

The third term in (1.35) can be simplified by introducing the following abbreviations

$$\begin{aligned} QM_{xx} &= \sum_i Q_i\left(3x_i^2-r_i^2\right), \\ QM_{yy} &= \sum_i Q_i\left(3y_i^2-r_i^2\right), \\ QM_{zz} &= \sum_i Q_i\left(3z_i^2-r_i^2\right), \\ QM_{xy} &= QM_{yx} = 3\sum_i Q_ix_iy_i, \\ QM_{xz} &= QM_{zx} = 3\sum_i Q_ix_iz_i, \\ QM_{yz} &= QM_{zy} = 3\sum_i Q_iy_iz_i \end{aligned} \tag{1.36}$$

Like the moment of inertia (Vol. 1, Chap. 5) that describes the mass distribution of a rigid body here we describe the spatial charge distribution by the components QM_{jk} of the quadrupole tensor

$$QM = \begin{pmatrix} QM_{xx} & QM_{xy} & QM_{xz} \\ QM_{yx} & QM_{yy} & QM_{yz} \\ QM_{zx} & QM_{zy} & QM_{zz} \end{pmatrix} \tag{1.37}$$

With this definition the third term in (1.35) (quadrupole term) becomes

$$\begin{aligned} \phi_Q = \frac{1}{8\pi\varepsilon_0 R^5}\big[&QM_{xx}X^2+QM_{yy}Y^2 \\ &+QM_{zz}Z^2+2\big(QM_{xy}XY \\ &+QM_{xz}XZ+QM_{yz}YZ\big)\big]. \end{aligned} \tag{1.38}$$

From (1.36) it follows that the quadrupole term is symmetrical and its trace (the sum of the main diagonal elements) is zero.

The quadrupole moment QM is a measure for the deviation of the charge distribution from spherical symmetry. A homogeneously charged sphere has $QM = 0$

Examples

With the charge distribution of Fig. 1.29 we get from (1.36)

$$\begin{aligned} &QM_{xx} = QM_{yy} = QM_{zz} = QM_{xz} = QM_{yz} = 0, \\ &QM_{xy} = 3\cdot a\cdot d\cdot Q. \end{aligned}$$

1.5 Conductors in an Electric Field

In an electric field the charges inside a conductor are subjected to forces $F = q \cdot E$. The forces move the charges inside the conductor as long as the charges on their new positions generate an electric field of equal magnitude but with direction opposite to the external field which is then compensated (Fig. 1.30). This displacement of charges in conductors is called **influence**.

Inside a conductor is no electric field. The charges are located only at the surface of the conductor.

1.5.1 Influence

Influence can be demonstrated by a simple experiment (Fig. 1.31).

We hold two metallic plates in contact into the electric field of a parallel plate capacitor Their handles are insulated from all conductors. Now we separate these plates inside the field and pull them out without any further contact. While the plates were in contact inside the field the charges moved to the outer sides of the conductors because of influence and after separation the plates had a surplus of charges $+Q$ resp. $-Q$. That can be proved by an electrometer.

Influence can be also impressively demonstrated by the ***beaker-electroscope*** (Faraday's ice-pail experiment), Fig. 1.32. We dip an insulated but electrically positive charged ball in a metallic beaker without touching the walls. The electric field exerts a force upon the free (negative) electrons and moves them to the inner side of the metallic beaker so that the electroscope sees a deficit of negative charges. That means that it is positively charged and shows a corresponding deflection. This deflection vanishes if we remove the ball inside the beaker (Fig. 1.32).

But, if the ball touches the inner wall of the beaker the ball discharges. Because of their repulsion the free charges move to the outside of the beaker while its interior remains free of electric fields. Now we can charge the ball again and discharge at the inner wall of the beaker. In this way the electroscope can be charged up to an arbitrary voltage (path one in Fig. 1.33). Its limit depends on the loss of charges by insufficient insulation of the electroscope. Using the way 2

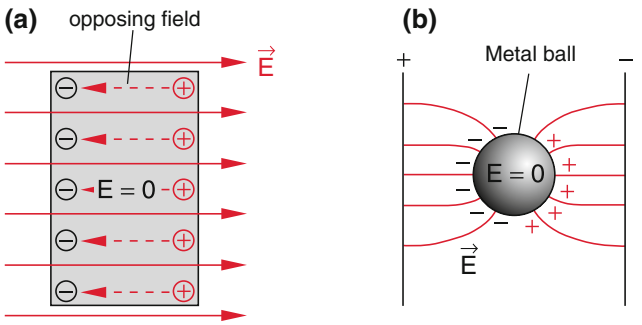


Fig. 1.30 Illustration of influence. **a**) For a conductor inside a parallel plate capacitor **b**) for a conducting sphere

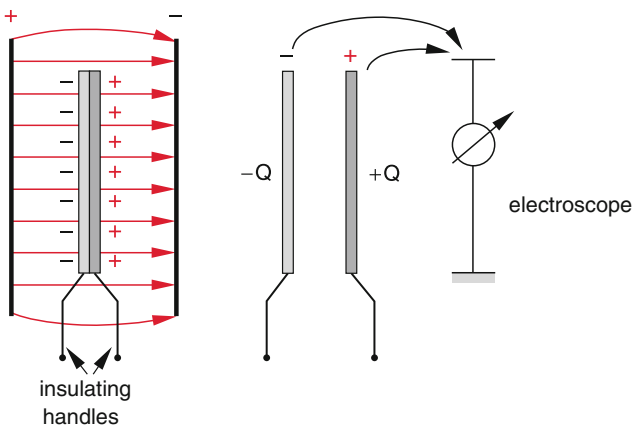


Fig. 1.31 Demonstration of influence

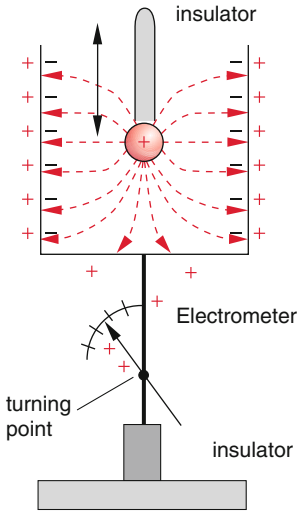


Fig. 1.32 Demonstration of influence when a charged ball is placed inside a conducting cup

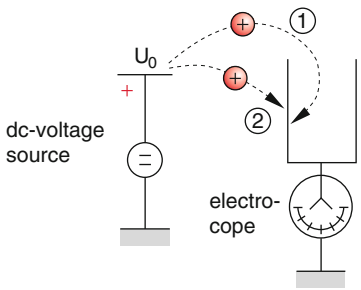


Fig. 1.33 Different achievable voltages when charging a conducting cup inside or outside the cup walls

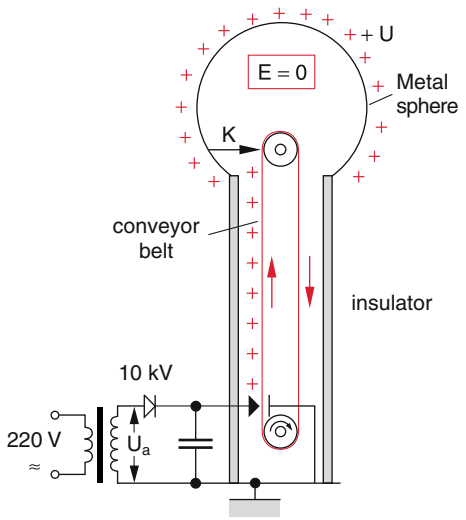


Fig. 1.34 Van de Graaff generator

where the transfer of charges is accomplished at the outer side of the beaker only a voltage not higher than that of the charge source U is possible.

The fact that we can transfer nearly any amount of charges to the inner walls of a spherical conductor and thus reach high voltages is used in the ***van de Graaff generator*** (Fig. 1.34).

An isolating belt moves about two cylinders one near the bottom and the other inside the top sphere of the generator. Because of the high field at the tips of a metallic comb charges generated by a DC supply are transferred to the belt. They are transported to the inner walls of the metallic sphere at the top. There another metallic comb that is connected to the inner wall of the sphere removes the charges from the belt. By influence charges are forced to the outside of the sphere and its interior is always free from electric fields. Even with demonstration devices we can generate a voltage of about 10^5 V that is limited only by coronary losses.

To avoid discharges and to reduce losses the high voltage generator is placed completely inside a housing that is filled with gas at higher pressure so that voltages above 1 MV are attained [4].

The fact that the space inside a closed surface of a conductor is free from electric fields is used by Faraday's cage (Fig. 1.35). To protect sensitive devices from high electric fields, high tension or lightning flashes one puts them into a grounded cage of conducting material.

1.5.2 Capacitors

A device consisting of two oppositely charged surfaces of conductors is called a capacitor. When we bring the charge $+Q$ on one of the two surfaces then by influence on the other surface which was initially neutral a separation of charges takes place. At the surface next to the charged plate a charge $-Q$ will arise while on the distant surface the charge $+Q$ will appear. If we connect the initially uncharged plate with the grounding point of the charging device for the first plate the charge $+Q$ drains away. The charge $-Q$ remains on the second plate (Fig. 1.36).

Because the electric field in the space between the conductors is proportional to the charge Q on the conductor also the voltage $U = \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s}$ is proportional to Q and we obtain the relation

$$Q = C \cdot U. \tag{1.39}$$

The constant of proportionality C is called the **capacitance** of the capacitor. It has the unit

$$[C] = 1\,\frac{\text{Coulomb}}{\text{Volt}} \overset{\text{def}}{=} \text{Farad} = 1\ \text{F}. \tag{1.40}$$

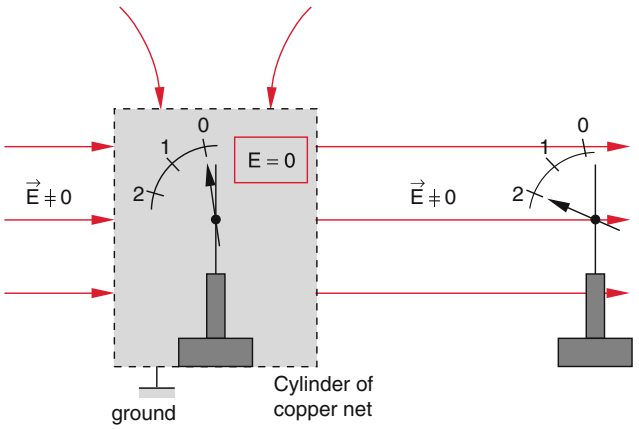


Fig. 1.35 Faraday's cage. The volume inside a grounded metallic surrounding is field-free

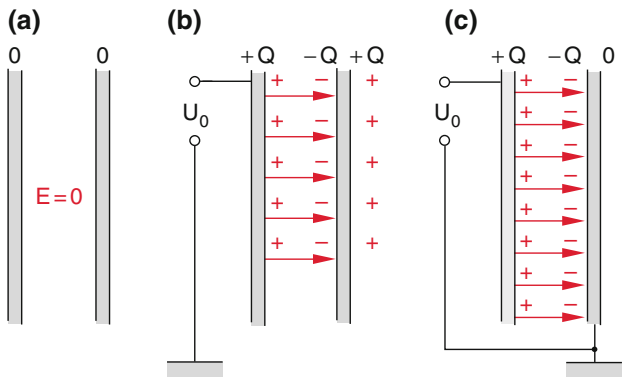


Fig. 1.36 Influence in a plane plate capacitor. **a**) Uncharged capacitor **b**) charging the left plate with the charge $+Q$. **c**) Grounding the right plate

Because 1 F is a very large capacity subunits are used

$$1\ \text{Picofarad} = 1\ \text{pF} = 10^{-12}\,\text{F},$$
$$1\ \text{Nanofarad} = 1\ \text{nF} = 10^{-9}\,\text{F},$$
$$1\ \text{Microfarad} = 1\ \mu\text{F} = 10^{-6}\,\text{F}.$$

Now we will calculate capacitance and field distribution of the most important types of capacitors and thereby illustrate the practical application of the Laplace equation.

1.5.2.1 Parallel Plate Capacitor

The plates of a parallel plate capacitor with the plates at $x = 0$ and $x = d$ (see Fig. 1.37) have the charges $+Q$ resp. $-Q$. The space between the plates is free of charges and so we can apply the Laplace equation (1.16b) in its one dimensional form

$$\frac{\partial^2\phi}{\partial x^2} = 0 \Rightarrow \phi = ax + b. \tag{1.41}$$

Let ϕ_1 be the potential of the left plate at $x = 0$ and ϕ_2 the potential of the right plate at $x = d$. Then the voltage between the plates is $U = \phi_1 - \phi_2$. From (1.41) follows

$$\phi_1 = b \quad \text{and} \quad \phi_2 = a \cdot d + \phi_1$$
$$\Rightarrow a = \frac{\phi_2 - \phi_1}{d} = -\frac{U}{d}.$$

The potential between the plates

$$\phi(x) = -\frac{U}{d} \cdot x + \phi_1 \tag{1.41a}$$

decreases linearly with the voltage (Fig. 1.37).

The electric field strength is

$$\boldsymbol{E} = -\mathbf{grad}\,\phi = \frac{U}{d} \cdot \hat{\boldsymbol{x}}. \tag{1.42}$$

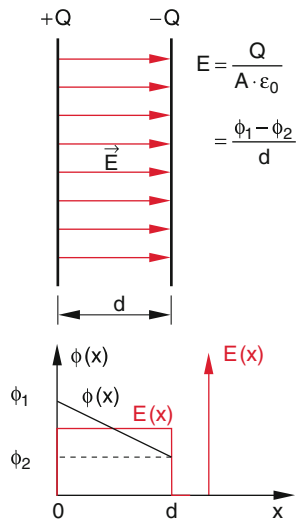


Fig. 1.37 Field distribution inside a plane plate capacitor

The magnitude of the field (1.42) is

$$E = \frac{U}{d}. \tag{1.42a}$$

Because for an area A of the plate the field strength is $E = Q/(A \cdot \varepsilon_0)$ (see (1.8b)) it therefore follows for the capacitance $C = Q/U$

$$C = \varepsilon_0 \cdot \frac{A}{d}. \tag{1.43}$$

The capacitance of a parallel plate capacitor is proportional to the area of the plate A and inversely proportional to the distance d between the plates.

Example

$$A = 100\ \text{cm}^2,\ d = 1\ \text{mm} \Rightarrow C = 88.5\ \text{pF}.$$

1.5.2.2 Spherical Capacitor

The spherical capacitor consists of two concentric spheres of radius $r_1 = a$ and $r_2 = b$ that carry the charges $+Q$ resp. $-Q$. (Fig. 1.38).

With the knowledge of Sect. 1.3.4 we can obtain field strength $E(r)$ and potential $\phi(r)$ immediately.

Inside ($r < a$) there is no field and the potential is constant. Because the function $E(r)$ must be finite, ϕ is continuous at $r = a$ and its value at $r \leq a$ is

$$\phi_1 = \frac{Q}{4\pi\varepsilon_0 a}. \tag{1.44a}$$

In the space between the two spheres ($a < r < b$) we have the field of a point charge at the center of the inner sphere

$$\boldsymbol{E}_2 = \frac{Q}{4\pi\varepsilon_0 r^2}\hat{\boldsymbol{r}} \tag{1.44b}$$

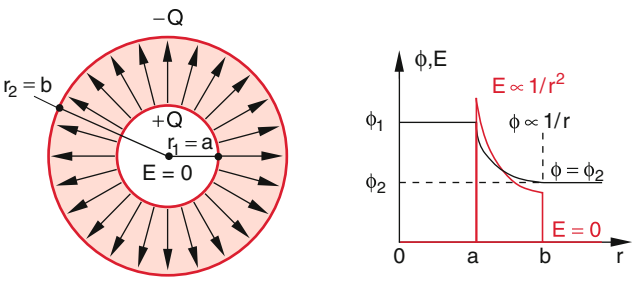


Fig. 1.38 Electric field and potential of a spherical capacitor

and the potential

$$\phi_2 = \frac{Q}{4\pi\varepsilon_0 r}. \tag{1.44c}$$

Outside of the outer sphere we have because of $Q_{\text{total}} = 0$ no field and a potential

$$\phi_3 = \frac{Q}{4\pi\varepsilon_0 b}. \tag{1.44d}$$

The voltage between the two spheres is

$$\begin{aligned} U = \phi_i - \phi_2 &= \frac{Q}{4\pi\varepsilon_0}\left(\frac{1}{a} - \frac{1}{b}\right) \\ &= \frac{Q}{4\pi\varepsilon_0}\frac{b-a}{ab}. \end{aligned} \tag{1.45}$$

Figure 1.38 shows the variation of $\phi(r)$ and $E(r)$ in the different ranges. At the charged surfaces of the conductors $E(r)$ has a discontinuity of $\Delta E = \sigma/\varepsilon_0$.

The capacitance of a spherical capacitor is according to (1.39) and (1.45)

$$C = \frac{Q}{U} = \frac{Q}{\phi_i - \phi_a} = \frac{4\pi\varepsilon_0 \cdot a \cdot b}{b-a}. \tag{1.46}$$

If the distance $d = b - a$ is small compared to a we get from (1.46) and the geometrically means $\overline{R} = (a \cdot b)^{1/2}$

$$C = \frac{4\pi\varepsilon_0 \overline{R}^2}{d} = \frac{\varepsilon_0 \cdot A}{d}, \tag{1.46a}$$

an expression similar to the flat parallel plate capacitor where A is now the area of a fictive sphere between the two conductors of the capacitor.

For the limit of $b \to \infty$ of the outer radius b we get from (1.46) the capacitance of a sphere of radius a and its second surface at infinity

$$C = 4\pi\varepsilon_0 \cdot a. \tag{1.46b}$$

Charging the sphere to a voltage U we get the charge

$$Q = 4\pi\varepsilon_0 a \cdot U. \tag{1.46c}$$

The capacitance of a conducting sphere is proportional to the radius of the sphere but not to its surface.

1.5.2.3 Capacitors in Parallel and in Series

Note A general note on drawings of electronic circuits: Drawings of electronic circuits consist of symbols for passive elements e.g. resistor, capacitor, inductor and active elements e.g. voltage source, transistor and so on. These elements are connected by lines that have no resistance! This means that at both ends of the line is the same potential and the same voltage.

Connecting several capacitors in parallel (Fig. 1.39a) the voltage at each capacitor is the same. Otherwise charges would move until the voltages are equal. The charges add and we have according to (1.39) for the total capacitance

$$C = \sum_i C_i. \tag{1.47}$$

When connecting several capacitors in series the charges are separated by influence so that two neighboring plates that are connected by a conductor have the same amount of charge but of opposed sign (Fig. 1.39b).

The voltage between the two plates that are connected by a conductor, is of course zero, because the voltage due to the external source is just compensated by the electric field generated by the two charges $+Q$ and $-Q$. For the total capacitance C we now get

$$\frac{1}{C} = \sum_i \frac{1}{C_i}. \tag{1.48}$$

For capacitors in series the total capacitance becomes smaller but the break down voltage increases.

We also can see this fact from the relation $U = \int \boldsymbol{E}\,\mathrm{d}\boldsymbol{s}$. For equal field strength $\boldsymbol{E}$ in each capacitor the voltages add,

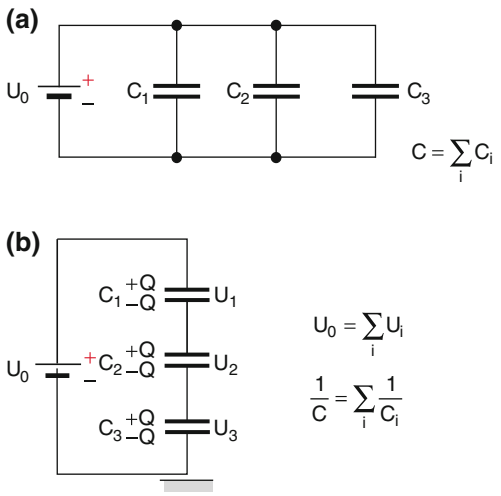


Fig. 1.39 Parallel and Series circuits of several capacitors

if the capacitors are connect in series. For the constant total charge Q we get because of $U = \sum U_i = \sum \frac{Q}{C_i} = Q/C$ the relation $1/C = \sum \frac{1}{C_i}$.

For parallel plate capacitors Eqs. (1.47) and (1.48) also can be derived from (1.43):

While for parallel connection the areas are added for series connection the distances add.

Example

The total capacitance of two capacitors is for parallel connection

$$C = C_1 + C_2$$

whereas for capacitors in series is

$$C = \frac{C_1 \cdot C_2}{C_1 + C_2}$$

To realize very large capacitances the area A of the conductors must be large and the distance between them as small as possible.

The practical realization uses wound capacitors. Two conducting foils separated by a thin isolating foil are wound to a cylinder. This gives a large area and a very small distance d. Figure 1.40b shows some commercial examples.

Often a variable capacitance is needed and can be realized by rotary capacitors (Fig. 1.40).

1.6 The Energy of the Electric Field

An isolated metallic sphere of radius a can be charged by transferring small charges $q = \mathrm{d}Q$ step by step e.g. by a "charge spoon". Each step requires the work

$$\begin{aligned} \mathrm{d}W &= \mathrm{d}Q \cdot (\phi_a - \phi_\infty) \\ &= \mathrm{d}Q \cdot \phi_a \quad \text{for} \quad \phi_\infty = 0 \end{aligned}$$

where according to (1.44a–1.44d)

$$\phi_a = \frac{Q}{4\pi\varepsilon_0 \cdot a}$$

is the potential of the sphere with charge Q (Fig. 1.41). To charge up to the total charge Q the work

$$W = \frac{1}{4\pi\varepsilon_0 \cdot a} \int Q \cdot \mathrm{d}Q = \frac{Q^2}{8\pi\varepsilon_0 \cdot a} = \frac{1}{2}\frac{Q^2}{C}$$

is needed where according to (1.46b) $C = 4\pi\varepsilon_0 \cdot a$ is the capacitance of the sphere.

The energy of the sphere that is charged to a voltage U against its surroundings is then

$$W = \frac{1}{2}\frac{Q^2}{C} = \frac{1}{2} \cdot C \cdot U^2, \quad \text{because} \quad Q = C \cdot U.$$

This result that we have derived for a charged sphere is valid generally for arbitrary capacitors (see Problem 1.12).

A capacitor of capacitance C charged up to a voltage U contains an energy

$$W = \frac{1}{2} C \cdot U^2, \tag{1.49}$$

that is stored as the energy of the electrostatic field.

The capacitance of the flat parallel plate capacitor of area A and plate separation d is $C = \varepsilon_0 \cdot A/d$ and its voltage is $U = E \cdot d$. Therefore the energy becomes

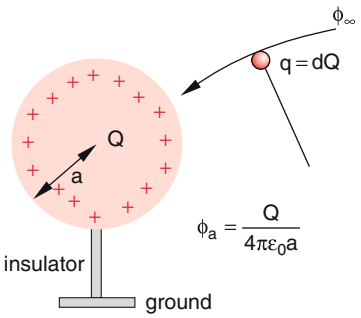


Fig. 1.41 Derivation of the potential and the charging energy of a conducting sphere

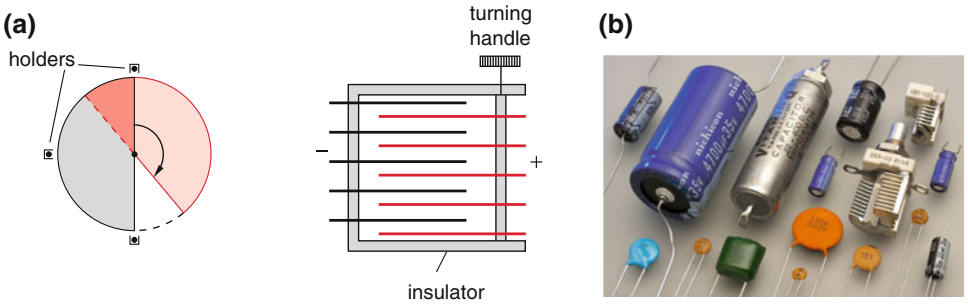


Fig. 1.40 **a**) Variable capacitor, **b**) different designs of commercial capacitors

$$W_{\mathrm{el}} = \frac{1}{2}\varepsilon_0 E^2 \cdot A \cdot d = \frac{1}{2}\varepsilon_0 E^2 \cdot V$$

The energy density of the electric field inside the capacitor is

$$w_{\mathrm{el}} = \frac{W_{\mathrm{el}}}{V} = \frac{1}{2}\varepsilon_0 \cdot E^2 \tag{1.50}$$

This result is valid for any electric field independent of its mode of generation.

Example

A capacitor $C = 1\ \mu\mathrm{F}$ which is charged to $U = 1$ kV stores the energy of $W = \frac{1}{2}CU^2 = 0.5\,\mathrm{J}$.

1. In great plasma fusion plants capacitor banks of $C = 0.1\,\mathrm{F}$ and $U = 50$ kV are used. Their stored energy is 125 MJ. If they are discharged in one millisecond the delivered electric power is $P = C \cdot U^2/10^{-3}\mathrm{W} = 1.25 \times 10^{11}\,\mathrm{W}$!
2. If we describe the electron by the model of a homogeneously charged sphere of radius r_e itselectrostatic energy is Wei $= e^2/8\pi\varepsilon_0 \cdot r_e$.
3. If we assume that this energy equals the rest energy $E = mc^2$ of the electron (see Vol. 1, Chap. 4) we get with the known value of the electron charge $e = 1.6 \times 10^{-19}\,\mathrm{C}$ and its mass $m_e = 9.108 \times 10^{-31}\,\mathrm{kg}$ the so called *classical electron radius* $r_e = 1.4 \times 10^{-15}\,\mathrm{m}$. Experiments show, however, (see Vol. 3) that the "real radius" of the electron must be much smaller. The simple model of a homogeneously charged sphere of radius r_e cannot be correct.

1.7 Dielectrics in Electric Fields

If we fill a parallel plate capacitor that holds a charge $Q = C \cdot U$ completely with a slab of dielectric material the voltage between the conducting plates decreases by a factor of ε. Because the charge Q is constant the capacitance C must have increased by a factor ε. Instead of (1.43) we get for the capacitance of a parallel plate capacitor

$$C_{\mathrm{Diel}} = \varepsilon \cdot C_{\mathrm{Vac}} = \varepsilon \cdot \varepsilon_0 \frac{A}{d} \quad \text{with} \quad \varepsilon > 1. \tag{1.51}$$

This number ε is named **relative dielectric constant** of the insulator. Such isolating materials are also called dielectrics. Table 1.1 lists the values ε of a few materials.

Table 1.1 Relative dielectric constants for some materials (Stöcker: Taschenbuch der Physik (Harri Deurtsch Frankfurt)

Material	ε_r
Quartzglass	3.75
Pyrexglas	4.3
Porcelain	6–7
Copper-oxyd	18
Cearamic	
TiO_2	≈80
$CaTiO_3$	≈160
$SrBi(TiO_3)$	≈1000
Liquids	
Water	81
Ethylalcohol	25.8
Benzene	2.3
Nitrobenzene	37
Gases	
Air	1.000576
H_2	1.000264
SO_2	1.0099

The magnitude of the electric field $|\boldsymbol{E}|$ is proportional to the voltage U and therefore E sinks too by the factor ε. For example, the field of a point charge Q inside a homogeneous insulator is

$$\boldsymbol{E} = \frac{1}{4\pi\varepsilon\varepsilon_0}\frac{Q}{r^2}\hat{\boldsymbol{r}}. \tag{1.52}$$

What is the reason for the reduction of the electric field?

1.7.1 Dielectric Polarization

Analogue to the phenomenon of influence the charges in the dielectric material are displaced in an external electric field. But, in an insulator the carriers of electric charges cannot move freely contrary to the situation in an electric conductor. Therefore they cannot move to the boundaries of the insulator. In an external electric field the charges can only be displaced inside the atoms or molecules of the insulator (Fig. 1.42).

For an atom in an external electric field the center of negative charges S^-, (electrons), and the center of positive charges S^+, (atomic nucleus), no longer coincide. The atoms have now become electric dipoles (Fig. 1.43).

These dipoles generated by an external electric field are called "induced dipoles".

If d is the distance between the centers of positive and negative charges then the induced dipole moment of each atom is

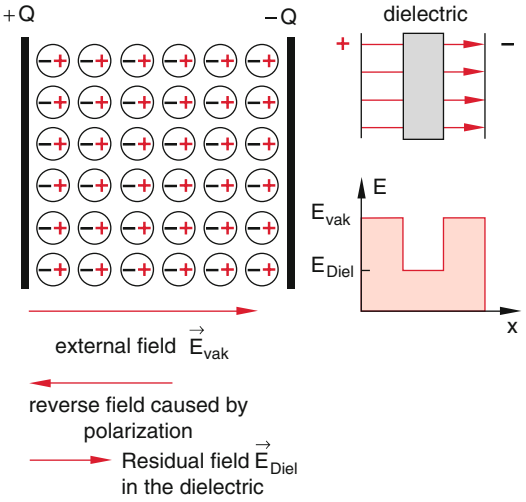


Fig. 1.42 Dielectric material inside a capacitor

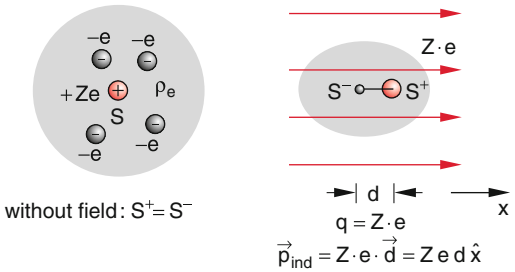


Fig. 1.43 Displacement of electric charges and their center points in an external electric field

$$\boldsymbol{p} = q \cdot \boldsymbol{d}.$$

The vector sum of the dipoles of all N atoms in the unit volume is the polarization

$$\boldsymbol{P} = \frac{1}{V}\sum_i \boldsymbol{p}_i. \tag{1.53a}$$

If we neglect all other interactions (e.g. thermal) all dipoles are aligned parallel to the electric field. Then, in a homogeneous field $\boldsymbol{E}$ the amount of the polarization becomes

$$P = N \cdot q \cdot d = N \cdot p, \tag{1.53b}$$

where N is the number of dipoles per unit volume. The displacement d between the centers of charge is determined by the condition that the restoring force of the attracting charges just compensates the external force $\boldsymbol{F} = q \cdot \boldsymbol{E}$.

In general the displacements are small compared to the diameter of the atom.

Because at small displacements the restoring force $-\boldsymbol{F}$ is proportional to the displacement d (Hooke's law) we get $d \propto E$. Therefore the dipole moment p in electric fields $E \leq 10^5$ V/cm is

$$\boldsymbol{p} = \alpha \cdot \boldsymbol{E} \tag{1.54}$$

The proportionality constant α is called atomic **polarizability**. It depends on the data of the atom and is a measure of the restoring forces that arise due to the displacement of charges. In general, α is a tensor i.e. p depends on the orientation in space

Example

The Na-atom has an atomic polarizability $\alpha = 3 \times 10^{-39}\,\mathrm{As\,m^2V^{-1}}$. In a field $E = 10^5$ V/m is $d = 1.88 \times 10^{-5}$ Å $= 1.88 \times 10^{-15}$ m.

In electric fields of technical applications is the displacement of charges very small compared to the diameter of atoms.

1.7.2 Polarization Charges

Because of the displacement of charges in an electric field at the front side of the dielectric material charges Q arise (Fig. 1.44) that are called polarization charges. Their surface charge density

$$\sigma_{\mathrm{pole}} = \frac{Q_{\mathrm{pole}}}{A} = \frac{N \cdot q \cdot d \cdot A}{A} = P \tag{1.55}$$

is equal to the amount of the polarization $\boldsymbol{P}$.

Inside the dielectric the negative and the positive charges compensate and there the total charge density is zero. These surface charges are opposite to the charges at the surface of the capacitor plates that are called free charges because they can move freely.

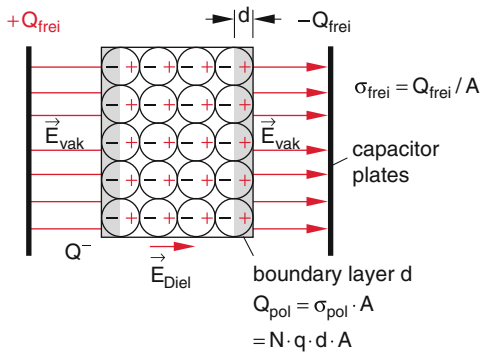


Fig. 1.44 Polarized dielectric inside a capacitor

In the homogeneous field $\boldsymbol{E}$ of the parallel plate capacitor without a dielectric it follows from the electric flux through an area A parallel to the plates (see Sect. 1.5)

$$\begin{aligned} \Phi_{\text{el}} &= \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} = \frac{Q}{\varepsilon_0} \\ \Rightarrow E \cdot A &= \frac{Q}{\varepsilon_0} \Rightarrow E = \frac{\sigma_{\text{free}}}{\varepsilon_0}. \end{aligned} \tag{1.56}$$

Inside the dielectric the external field $E = \sigma_{\text{free}}/\varepsilon_0$ and the opposite field due to polarization $E_{\text{pole}} = \sigma_{\text{pole}}/\varepsilon_0$ superimpose and the resulting field in the dielectric becomes

$$\boldsymbol{E}_{\text{Diel}} = \frac{\sigma_{\text{free}} - \sigma_{\text{pole}}}{\varepsilon_0}\hat{\boldsymbol{e}} = \boldsymbol{E}_{\text{Vac}} - \frac{\boldsymbol{P}}{\varepsilon_0} \tag{1.57}$$

Because of $\boldsymbol{P} \| \boldsymbol{E}$ it follows.

The field inside the dielectric becomes lower.

Note For the induced dipoles is $\boldsymbol{p} = \alpha \cdot \boldsymbol{E}$. Therefore the direction of the polarization P inside the dielectric is the same as that of the external electric field creating the induced dipoles. The direction of the electric dipole moment $\boldsymbol{p}$ is defined as the direction from the negative to the positive charge of the dipole while the electric field lines are directed from the positive to the negative charge. Therefore the field inside the dielectric generated by dipoles is opposite to the external electric field and reduces it. The amount of the electric field generated by the dipoles is smaller than the external field so the remaining field inside the dielectric has the same direction as the external field but has a smaller amount.

With (1.53a, 1.53b) and (1.54) the polarization P can be written as

$$\boldsymbol{P} = N \cdot \boldsymbol{E}_{\text{Diel}}. \tag{1.58}$$

Now we introduce the dielectric susceptibility $\chi = N \cdot \alpha/\varepsilon_0$ and we get from (1.57) and (1.58)

$$\boldsymbol{P} = \varepsilon_0 \chi \boldsymbol{E}_{\text{Diel}} \quad \text{and} \quad \boldsymbol{E}_{\text{Diel}} = \frac{\mathbf{E}_{\text{Vac}}}{1+\chi}. \tag{1.59}$$

The comparison with

$$E_{\text{Diel}} = \frac{1}{\varepsilon} E_{\text{Vac}}$$

yields the relative dielectric constant

$$\varepsilon = 1 + \chi = 1 + (N \cdot \alpha/\varepsilon_0)$$

and the polarization

$$\boldsymbol{P} = \varepsilon_0(\boldsymbol{E}_{\text{Vac}} - \boldsymbol{E}_{\text{Diel}}).$$

Influence and polarization are principally equal phenomena. They describe the displacement of charges in matter due to an external electric field. In conductors the charges can move freely up to the surface. The field inside the conductor is completely compensated (influence).

In insulators charges can only be displaced inside the atoms (polarization). Surface charges are generated. The field inside is only partially compensated (1.57). Field strength and voltage are lowered by the factor ε. The capacitance of a capacitor with dielectric rises correspondingly by the factor ε.

When we bring a conducting slab of thickness b between the plates of a flat plate capacitor that is charged to a voltage U_0, of area A and distance d then the voltage will sink from $U_0 = Q \cdot d/(\varepsilon_0 A)$ to $U = Q/(\varepsilon_0 A) \cdot (d - b)$ and the capacitance rises correspondingly to

$$C = \frac{Q}{U} = \frac{A\varepsilon_0}{d-b},$$

because the effective distance between the plates is now only $d - b$.

With a dielectric of thickness $b < d$ C rises to

$$C = \frac{A_0\varepsilon_0}{d - b(\varepsilon - 1)/\varepsilon}.$$

1.7.3 Equations of the Electrostatic Field in Matter

In a homogeneous field the positive and negative polarization charges compensate inside the dielectric and only at the surface of the dielectric are non-compensated polarization charges of one polarity. Therefore at a boundary of the dielectric perpendicular to the external field E a discontinuity of the field strength from E_{vac} to E_{diel} must exist.

In an inhomogeneous field the polarization P is not the same at every position. Now we have polarization charges also in the interior of the dielectric. Because of the spatially varying displacement of charges there are no more the same number of opposite charges in a volume element that can compensate each other.

Consider a volume V with a surplus of charges ΔQ_{pol} due to the spatially varying polarization (Fig. 1.45). We describe these charges by a spatially varying polarization charge density ΔQ_{pol} which has been taken away from volume V by displacement of charges through the surface A that encloses the volume

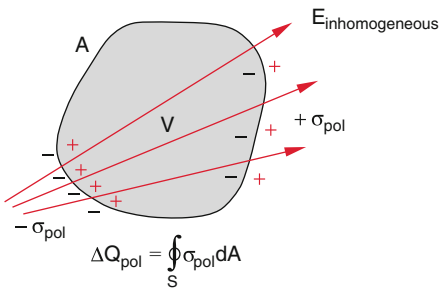


Fig. 1.45 Polarization of a dielectric material in an inhomogeneous field

$$\Delta Q_{\text{pol}} = -\int_V \varrho_{\text{pol}}\,\mathrm{d}V.$$

This is according to (1.55)

$$\Delta Q_{\text{pol}} = \int_A \sigma_{\text{pol}}\,\mathrm{d}\boldsymbol{A} = \int_A \boldsymbol{P}\cdot\mathrm{d}\boldsymbol{A}. \tag{1.60a}$$

Using Gauss' law we convert the surface integral into a volume integral

$$\int_A \boldsymbol{P}\cdot\mathrm{d}\boldsymbol{A} = \int_V \operatorname{div}\boldsymbol{P}\,\mathrm{d}V = -\int \varrho_{\text{pol}}\,\mathrm{d}V \tag{1.60b}$$

and obtain by comparison with (1.60a, 1.60b)

$$\operatorname{div}\boldsymbol{P} = -\varrho_{\text{pol}}, \tag{1.61}$$

In other words:

The polarization charges of density ϱ_{pol}, generated by the external electric field, are the sources of the electric polarization.

Equation (1.61) relates the polarization of matter to the spatial density of polarization charges and corresponds to the equation

$$\operatorname{div}\boldsymbol{E} = \varrho/\varepsilon_0$$

for free charges.

In matter the opposing polarization charges ϱ_{pol} are added to the free charges and the electric field $\boldsymbol{E}_{\text{Diel}}$ in the dielectric is

$$\operatorname{div}\boldsymbol{E}_{\text{Diel}} = \frac{1}{\varepsilon_0}\left(\varrho_{\text{free}} + \varrho_{\text{pol}}\right). \tag{1.62}$$

Because of $\boldsymbol{E}_{\text{Diel}} = \boldsymbol{E}_{\text{Vac}} - \boldsymbol{P}/\varepsilon_0$ we can rewrite (1.62) as

$$\operatorname{div}(\boldsymbol{E}_{\text{Vac}} - \boldsymbol{P}/\varepsilon_0) = \frac{1}{\varepsilon_0}\left(\varrho_{\text{free}} + \varrho_{\text{pol}}\right),$$

which with (1.10) again yields (1.61). With the ***dielectric displacement density***

$$\boldsymbol{D} \overset{\text{Def}}{=} \varepsilon_0\boldsymbol{E}_{\text{Diel}} + \boldsymbol{P} = \varepsilon\cdot\varepsilon_0\cdot\boldsymbol{E}_{\text{Diel}} = \varepsilon_0\cdot\boldsymbol{E}_{\text{Vac}} \tag{1.63}$$

the Poisson equation for the electric field can be written in generalized form

$$\operatorname{div}\boldsymbol{D} = \varrho, \tag{1.64a}$$

where $\varrho = \varrho_{\text{free}}$ is the original i.e. the free charge density of the considered volume. Equation (1.64a) is valid in matter but also in vacuum where $\varepsilon = 1$ and therefore $\boldsymbol{D} = \varepsilon_0\cdot\boldsymbol{E}$. In a space free from charges ($\varrho = 0$) is then

$$\operatorname{div}\boldsymbol{D} = 0. \tag{1.64b}$$

The unit of D is

$$[D] = [\varepsilon_0\cdot E] = 1\,\frac{\mathrm{A\,s}}{\mathrm{m}^2} = 1\,\frac{\mathrm{C}}{\mathrm{m}^2}.$$

D gives the surface charge density that has been displaced by the external field.

At a boundary between dielectric and vacuum the normal component of D is continuous because of

$$\varepsilon_0\cdot\varepsilon\cdot\boldsymbol{E}_{\text{diel}} = \varepsilon_0\cdot\boldsymbol{E}_{\text{vac}}$$

This is not valid for the tangential component of $\boldsymbol{D}$ as we will derive from fundamental properties of the electric field.

In Figs. 1.9, 1.10 and 1.11 we see that there are no closed electric field lines generated by charges. If there were closed field lines then a charge would move along these lines parallel to the field and would gain the energy $W = q\cdot\int\boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s}$ on each revolution. According to (1.50) the stored energy density, i.e. the energy per unit volume, $w = \frac{1}{2}\varepsilon_0 E^2$ would not be reduced by this action but the energy of the total system would increase. That contradicts the conservation of energy. Therefore it must be $W = 0$. Using Stokes' theorem (see Vol. 1, Sect. 8.6.1) the equation

$$\int \boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s} = 0 \tag{1.65a}$$

can be rewritten as

$$\int \mathbf{rot}\,\boldsymbol{E}\cdot\mathrm{d}\boldsymbol{A} = 0, \tag{1.65b}$$

where A is an arbitrary surface with the boundary line s. Equation (1.65a) is valid for each closed path and therefore (1.65b) is valid for each surface A. From this follows

$$\mathbf{rot}\, \boldsymbol{E} \equiv \mathbf{0}, \tag{1.65c}$$

This expresses the fact the static electric field $\boldsymbol{E}$ has no closed lines, it is "curl-free" and it preserves energy i.e. it is a conservative force field.

Note We can derive (1.65c) also from $\boldsymbol{E} = -\mathbf{grad}\, \phi$ and rot grad $\phi \equiv 0$ (Vol. 1, (13.26)). This means that $\boldsymbol{E}$ can be written as the gradient of a potential which is another statement of the fact that $\boldsymbol{E}$ is conservative.

In the next parts this knowledge will help us to understand the behavior of the electric field at a boundary between vacuum and dielectric. Of course, this can be also applied to a boundary of two dielectrics with their dielectric constants ε_1 and ε_2.

If the field vector E at the interface between vacuum and the dielectric is normal to the surface it is weakened by the factor ε according to (1.57). For a field that enters the dielectric under an angle α between field vector and normal (Fig. 1.46a) we separate the vector E at the surface into a component $\boldsymbol{E}_\perp$ normal and a component $\boldsymbol{E}_\parallel$ parallel to the boundary. We are now interested in the behavior of $\boldsymbol{E}_\parallel$ at the boundary. We consider the integration $\oint \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s}$ along the rectangular path ABCD of Fig. 1.46b. The thickness d of this rectangle is negligibly small so that practically only the path AB in vacuum and the path CD in the dielectric count.

Because of

$$\int_A^B \boldsymbol{E}_\parallel^{\mathrm{Vac}} \cdot \mathrm{d}\boldsymbol{s}_1 + \int_C^D \boldsymbol{E}_\parallel^{\mathrm{Diel}} \cdot \mathrm{d}\boldsymbol{s}_2 = \oint \mathbf{E} \cdot \mathrm{d}\boldsymbol{s} = 0$$

and $\mathrm{d}\boldsymbol{s}_1 = -\mathrm{d}\boldsymbol{s}_2$ it follows

$$\boldsymbol{E}_\parallel^{\mathrm{Vac}} = \boldsymbol{E}_\parallel^{\mathrm{Diel}} \tag{1.66a}$$

We get the following law of refraction for the electric field (Fig. 1.46a). If the $\boldsymbol{E}$-vector at the boundary between vacuum and dielectrics forms at the vacuum side the angle α to the boundary normal it has inside the dielectric the angle β against the normal of the boundary. Because $E_\perp^{\mathrm{Vac}} = \varepsilon \cdot E_\perp^{\mathrm{Diel}}$ we get

$$\tan\beta = \frac{E_\parallel^{\mathrm{Diel}}}{E_\perp^{\mathrm{Diel}}} = \varepsilon \cdot \frac{E_\parallel^{\mathrm{Vac}}}{E_\perp^{\mathrm{Vac}}} = \varepsilon \cdot \tan\alpha. \tag{1.66b}$$

From this we obtain with $\boldsymbol{D} = \varepsilon\varepsilon_0\boldsymbol{E}$

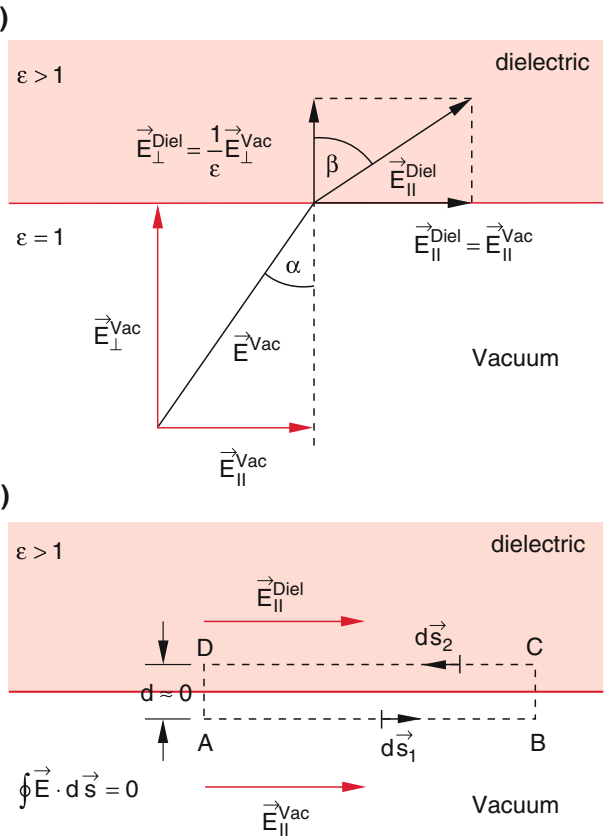


Fig. 1.46 **a**) Electric field at the boundary between dielectric and vacuum **b**) Illustration of the behavior of the tangential component at the boundary

$$\boldsymbol{D}_\parallel^{\mathrm{Vac}} = \frac{1}{\varepsilon}\boldsymbol{D}_\parallel^{\mathrm{Diel}}. \tag{1.66c}$$

This implies that the charge density ϱ is responsible for the fact that the field in the dielectric is larger than that of the free charge density generating the field in vacuum.

This can be illustrated by using a capacitor that is half filled with a dielectric (Fig. 1.47). In such an arrangement the charges on the plates are no more equally distributed but increase discontinuously at the border between vacuum and dielectric.

The capacitor of Fig. 1.47 is equivalent to the circuit of two parallel capacitors where one of them is filled with a dielectric. At equal voltages across both capacitors (parallel circuit!) the second capacitor contributes more free charges on its plates corresponding to the dielectric constant. Of course at equal voltages also the field $\boldsymbol{E}_\parallel$ parallel to the boundary between vacuum and dielectric is equal in both parts (1.66a) but $\boldsymbol{D}_\parallel$ is higher by the factor ε.

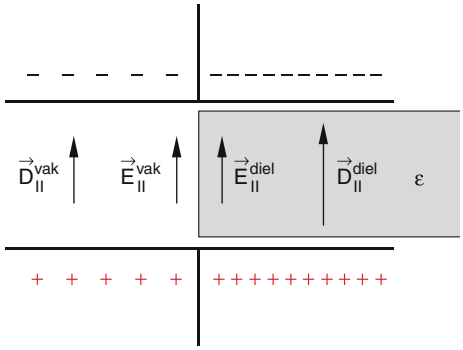


Fig. 1.47 Electric field E and dielectric displacement density D in an electric field without and with dielectric medium

If we fill a charged capacitor with a dielectric then the free charges move until (1.66a) resp. (1.66c) are satisfied.

1.7.4 The Electric Field Energy in Dielectrics

Consider a capacitor with the constant voltage U across its terminals. Filling the volume between the plates of this capacitor with a dielectric the initial capacitance C rises by the factor ε and so the charge density. The stored electric energy is

$$W_{el} = \frac{1}{2}CU^2 = \frac{1}{2}\varepsilon \cdot \varepsilon_0 \frac{A}{d}(d \cdot E)^2$$
$$= \varepsilon \cdot \frac{1}{2}\varepsilon_0 E^2 \cdot A \cdot d = \varepsilon \cdot \frac{1}{2}\varepsilon_0 \cdot E^2 \cdot V$$

and the energy density $w_{el} = W_{el}/V$ with $D = \varepsilon\varepsilon_0 E$ becomes

$$w_{el} = \varepsilon \cdot \frac{\varepsilon_0}{2}E^2 = \frac{1}{2}E \cdot D. \tag{1.67}$$

with $D = \varepsilon \cdot \varepsilon_0 \cdot E$

Equation (1.67) is the general form of (1.50) and is valid in vacuum ($D = \varepsilon_0 E$) and in matter.

One can understand the increase of energy density occurring when inserting a dielectric as follows: To the energy $\frac{1}{2}\varepsilon_0 E^2$ of the field in vacuum one has to add the energy necessary to displace the charges Q in the atoms by the distance x against the restoring force $\boldsymbol{F} = -k\boldsymbol{x} = Q \cdot \boldsymbol{E}$.

This additional energy is for one induced dipole

$$W_{pol} = -\int_0^d F\,dx = \frac{1}{2}kd^2 \quad \text{with} \quad k = \frac{Q \cdot E}{d}$$
$$\Rightarrow W_{pol} = \frac{1}{2}Q \cdot E \cdot d = \frac{1}{2}p \cdot E. \tag{1.68}$$

With Eqs. (1.60a, 1.60b) we get for N induced dipoles per unit volume the necessary energy density for polarizing the dielectrics that has to be added to the energy density $\frac{1}{2}\varepsilon_0 E^2$ of the field in vacuum. Altogether we get then the energy density (1.67).

$$w_{el} = \frac{1}{V}W_{pol} = \frac{1}{2}NpE = \frac{1}{2}P \cdot E$$
$$= \frac{1}{2}\varepsilon_0(\varepsilon - 1)E^2, \tag{1.69}$$

$$w_{el}^{diel} = \frac{1}{2}\varepsilon\varepsilon_0 E^2 = \frac{1}{2}ED \tag{1.70}$$

If the charged capacitor of Fig. 1.47 is switched off the voltage supply the capacitor remains charged but by inserting a dielectric the voltage does not remain constant. The energy stored in the field is reduced. Without the dielectric the energy is $W = \frac{1}{2}E_0 D_0 V$ where V is the volume of the capacitor between the plates. If the dielectric fills the volume completely then $D_1 = D_0$ (because of $\varrho_{tot} = \text{constant}$) and $E_1 = E_0/\varepsilon$. Therefore the energy $W = \frac{1}{2\varepsilon}E_0 D_0 V$ is smaller than that without dielectric.

A dielectric is pulled into an isolated charged capacitor. Mechanical energy is gained at the expense of electric energy. Work has to be done to remove the dielectric from the capacitor.

For an isolated capacitor it is easy to see that the dielectric is pulled into the capacitor. The system of capacitor-dielectric is closed and the energy released from the field is transferred to kinetic energy of the dielectric.

More difficult to understand is the case where the voltage across the capacitor is fixed (e.g. via a connection to a battery, Sect. 2.8). Inserting the dielectric into the capacitor causes a flow of charges from the battery onto the plates. The $\boldsymbol{D}$-field becomes larger by the factor ε and the energy rises (at constant $\boldsymbol{E}$-field) by the same factor ε.

Nevertheless the dielectric is pulled into the capacitor! Now the system capacitor-dielectric is no longer a closed system. The battery supplies charges and thus energy into the system capacitor-dielectric. If the dielectric is completely inserted then the surplus of charges on each plate is larger by the factor ε. To reach this goal work $W_{Bat} = \Delta Q \cdot U$ has to be supplied.

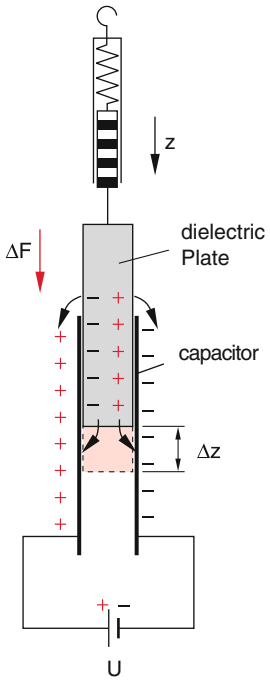


Fig. 1.48 A dielectric is pulled into a charged capacitor

Only half of this energy is used to increase the field energy. The rest is transformed into kinetic energy of the dielectric.

An experimental application of this fact is the determination of the dielectric constant ε of materials. A dielectric slab mounted to a spring balance is brought between the plates of an uncharged flat capacitor, so that only part of the capacitor is filled with material (Fig. 1.48).

Now we apply a voltage U across the capacitor and the slab is pulled by the distance Δz into the capacitor and the spring balance shows an additional force $\Delta F = k \cdot \Delta z$ *which* is caused by the attraction between the free charges on the capacitor and the induced surface charges of the dielectric. The work $\Delta W = \Delta F \cdot \Delta z$ which has to be done against the force of the spring is identical to the increase of the field energy

$$\begin{aligned} \Delta W_{\text{mech}} &= \Delta W_{\text{field}} = \frac{1}{2}(C_{\text{Diel}} - C_{\text{Vac}})U^2 \\ \Delta W &= \frac{1}{2}\varepsilon_0(\varepsilon - 1)b \cdot \Delta z U^2/d. \end{aligned} \tag{1.71a}$$

Because of $\Delta W = \Delta F \cdot \Delta z$ we get

$$\Delta F = \frac{1}{2}\varepsilon_0(\varepsilon - 1)b \cdot U^2/d \tag{1.71b}$$

which gives the value of ε.

In a second experiment we dip only a small part of the parallel plate capacitor into a dielectric liquid (e.g. nitrobenzene). The plates have a distance d and a width b. We apply a voltage U across the capacitor and the liquid

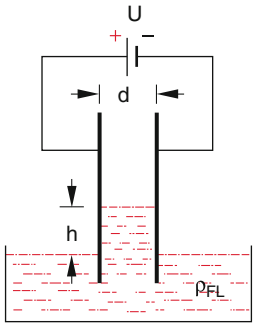


Fig. 1.49 A dielectric liquid is pulled into a charged capacitor until it has reached a height h

inside the capacitor rises by the height h above the surface of the liquid outside the capacitor (Fig. 1.49). The height h adjusts so that the mechanical work necessary to lift the volume of liquid in the vertical z-direction

$$W_{\text{mech}} = \int_{z=0}^{h} \varrho_{\text{Fl}} \cdot g \cdot b \cdot d \cdot z \, \mathrm{d}z = \frac{1}{2}\varrho_{\text{Fl}} \cdot g \cdot h \cdot V \tag{1.72a}$$

with $V = d \cdot b \cdot h$ is equal to the change in the field energy delivered by the battery

$$W_{\text{el}} = \frac{1}{2}\varepsilon_0(\varepsilon - 1)E^2 V \quad \text{with} \quad E = U/d \tag{1.72b}$$

Equating (1.72a) and (1.72b) yields the height

$$h = \frac{\varepsilon_0(\varepsilon - 1)}{\varrho_{\text{Fl}} \cdot g} E^2. \tag{1.73}$$

1.8 Atomic Fundamentals of Charges and Electric Moments

As mentioned in Sect. 1.1 the material carriers of charges are electrons with negative charge $-e$, and protons with positive charge $+e$. The first quantitative measurements of the charge of electrons has been made by Robert Andrews Millikan (1868–1953, Fig. 1.50a) in 1909 with his famous ***oil-drop experiment*** [5]. Because of its fundamental importance we will dis cuss it here in some more detail.

1.8.1 The Millikan Experiment

Spraying of oil produces small droplets that diffuse between the horizontal plates of a capacitor (Fig. 1.50). By friction during spraying the droplets become electrically charged with charges $q = n \cdot e$ where $n = 1, 2, 3, \ldots$.

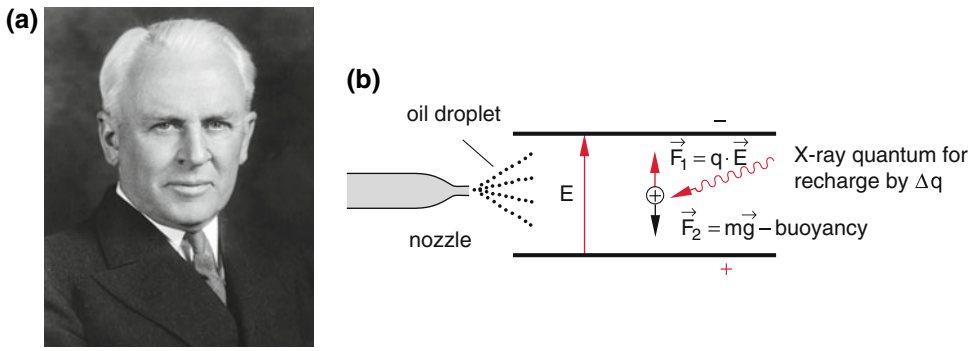


Fig. 1.50 **a**) Robert Andrews Millikan (Nobel Prize 1923) **b**) Experimental arrangement for the oil drop experiment [5]

In a capacitor without electric field the droplets of mass m and radius R descend with constant speed v if the force of gravity $F_g = m \cdot g$ just compensates the sum of the opposite lift forces $F_B = \varrho_{\text{Air}} \cdot \frac{4}{3}\pi \cdot R^3$ of buoyancy and $F_F = 6\pi\eta R \cdot v$ of Stokes friction (see Vol. 1, Sect. 8.5.4). The measurement of the constant velocity v of descent yields the radius R of the oil droplets

$$R = \left\{ \frac{9\eta \cdot v}{2g(\varrho_{\text{Oil}} - \varrho_{\text{Air}})} \right\}^{1/2}$$

and therefore also the mass $m = \frac{4}{3}\pi R^3 \varrho_{\text{Oil}}$ of the droplet since the density of oil is known.

Now we apply a suitable voltage U across the capacitor (distance of plates d). Now in the electric field $E = U/d$ the oil drops can be kept at an equilibrium position if the electric force $F_{\text{el}} = n \cdot e \cdot E$ acting on a droplet with n elementary charges e compensates the gravity force reduced by the buoyancy in air.

From this we get the charge

$$n \cdot e = (\varrho_{\text{Oil}} - \varrho_{\text{Air}})\, g \cdot \frac{4}{3}\pi R^3 / E. \tag{1.74}$$

To determine the integer number n we change the charge of the drops within the capacitor by irradiation with ionizing X-rays. This causes changes by small multiples Δn of the elementary charge e so that $\Delta q = \Delta n \cdot e$. Now the voltage has to be adjusted to keep the droplet again at an equilibrium position. From (1.74) follows for the equilibrium-voltages U_1 and U_2 before resp. after the alteration of charge

$$\frac{n_1 + \Delta n}{n_1} = \frac{U_1}{U_2} \Rightarrow \Delta n = -n_1 \frac{\Delta U}{U_2}. \tag{1.75}$$

The smallest change of charge is $\Delta n = 1\ldots$ Measurements with different values of Δn allow the determination of the discrete values Δn and n from the measurements of the difference $\Delta U = U_2 - U_1$. This gives from (1.74) the elementary charge e.

Today the accepted value of the elementary charge is

$$e = 1.602176487(40) \times 10^{-19}\,\text{C}$$

with a relative uncertainty of 2.5×10^{-8}

1.8.2 Deflection of Electrons and Ions in Electric Fields

If a particle of mass m and charge q is accelerated by the voltage U to the kinetic energy $E_{\text{kin}} = \frac{1}{2}mv^2 = q \cdot U$ i.e. to a speed of

$$v_0 = (2q \cdot U/m)^{1/2} \tag{1.76}$$

and then passes through a homogeneous electric field $\boldsymbol{E}$ (Fig. 1.51) it is deflected by the constant force $\boldsymbol{F} = q \cdot \boldsymbol{E}$. The trajectory of the particle becomes a parabola (compare it to the parabola of the horizontal throw in the gravity field).

For $\boldsymbol{v_0} = \{v_x, 0, 0\}$ and $\boldsymbol{E} = \{0, 0, E_z\}$ we get the deflection

$$\Delta z(x) = \frac{1}{2}at^2 = \frac{qE}{2m}\frac{x^2}{v_x^2}. \tag{1.77}$$

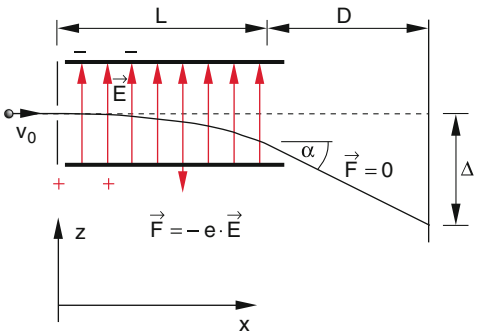


Fig. 1.51 Deflection of an electron in a homogeneous electric field

At the end of the capacitor ($x = L$) and with (1.76) we have

$$\Delta z(L) = \frac{E \cdot L^2}{4U}$$

The slope α of the trajectory is given by

$$\tan\alpha = \left(\frac{\mathrm{d}z}{\mathrm{d}x}\right)_{x=L} = \frac{qE}{m}\frac{L}{v_x^2} = \frac{E \cdot L}{2U}.$$

On the fluorescent screen at the distance D behind the end the capacitor the deviation

$$\begin{aligned}\Delta z(L+D) &= \frac{EL^2}{4U} + D \cdot \tan\alpha \\ &= \frac{EL}{2U}\left(\frac{L}{2} + D\right)\end{aligned} \qquad (1.78)$$

can be measured.

1.8.3 Molecular Dipole Moments

A molecule consists of K nuclei $(K = 2, 3, \ldots)$ with their positive charges $+Z_k \cdot e$ and of

$$Z_\mathrm{e} = \sum_{k=1}^{K} Z_k$$

electrons. The center of charges S^+ of the positive charges is chosen as the origin of the coordinate system. Then the center S^- of the electron charges with the coordinates $\boldsymbol{r}_i$ of the electrons is located at

$$\boldsymbol{d} = \frac{1}{Z_\mathrm{e}} \sum_{i=1}^{Z_\mathrm{e}} \boldsymbol{r}_i.$$

The dipole moment of the molecule is

$$\boldsymbol{p} = Q \cdot \boldsymbol{d} \quad \text{with} \quad Q = Z_\mathrm{e} \cdot e. \qquad (1.79)$$

The quantity d is the distance between positive and negative centers of charge (Fig. 1.52).

If both centers of charges coincide ($d = 0$) as for example for atoms or molecules with two equal atoms, the electric dipole moment becomes zero

However, in an electric field such non-polar molecules obtain an induced dipole moment because the centers of charge are displaced against each other (Fig. 1.43) In an inhomogeneous field all dipoles suffer a force $\boldsymbol{F} = \boldsymbol{p} \cdot \mathbf{grad}\, \boldsymbol{E}$.

An example for these forces is the attachment of neutral molecules on ions in an electrolyte (Fig. 1.53).

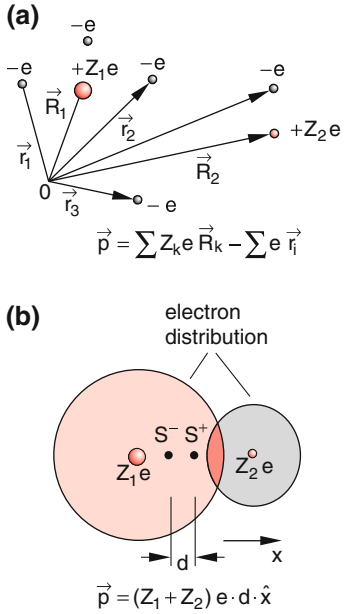


Fig. 1.52 **a**) Electric dipole moment of a molecule with $Z = Z_1 + Z_2$ electrons and an equal amount of positive charges of the nuclei. **b**) Displacement of the centers of positive and negative charges in an external electric field

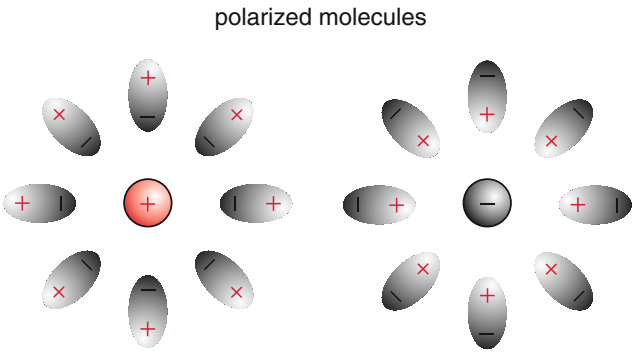


Fig. 1.53 Attraction of polar molecules by a positive charge

For most molecules which do not consist of equal atoms is $d \neq 0$. Such polar molecules therefore have non-vanishing electric dipole moments.

$$\boldsymbol{p} = Q \cdot \boldsymbol{d}.$$

Example

The molecule H_2O has a dipole moment $p = 6 \times 10^{-30}$ C m because the center of the negative charge $Q = -10e = -1.6 \times 10^{-18}$ C has a distance of about 4 pm to the positive center of charge (Fig. 1.54).

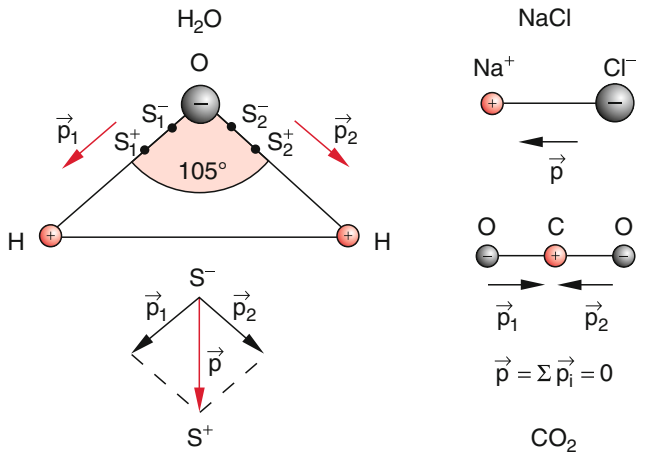


Fig. 1.54 Some examples of polar and nonpolar molecules

Table 1.2 List of some polar molecules

Molecule	Dipole moment/D
NaCl	9.00
CsCl	10.42
CsF	7.88
HCl	1.08
CO	0.11
H_2O	1.85
NH_3	1.47
C_2H_5OH	1.69

In molecular physics often the non-SI unit 1 *Debye* is used for the molecular dipole moment.

$$1\ \text{Debye} = 3.3356 \times 10^{-30}\,\text{C m}$$

Figure 1.54 shows a few examples of polar and non-polar molecules and Table 1.2 lists some values.

The potential energy of the interaction between two dipoles $\boldsymbol{p}_1$ and $\boldsymbol{p}_2$

$$W_{\text{pot}} = -\boldsymbol{p}_1 \cdot \boldsymbol{E}_2 = -\boldsymbol{p}_2 \cdot \boldsymbol{E}_1,$$

is obtained from (1.28) where $\boldsymbol{E}_i$ is the electric field of $\boldsymbol{p}_i$ at the position of dipole $\boldsymbol{p}_k$.

Using (1.25a) for the electric field $\boldsymbol{E}_i$ of dipole $\boldsymbol{p}_i$ yields with a distance $R \gg d = p/Q$ between the midpoints of both dipoles $\left(\widehat{\boldsymbol{R}} = \boldsymbol{R}/|\boldsymbol{R}|\right)$ at random orientations of both dipoles

$$\begin{aligned} W_{\text{pole}} &= \frac{1}{4\pi\varepsilon_0 R^3}\left[\boldsymbol{p}_1 \cdot \boldsymbol{p}_2 - 3\left(\boldsymbol{p}_1 \cdot \widehat{\boldsymbol{R}}\right)\left(\boldsymbol{p}_2 \cdot \widehat{\boldsymbol{R}}\right)\right] \\ &= \frac{p_1 p_2}{4\pi\varepsilon_0 R^3}\left[\cos(\boldsymbol{p}_1, \boldsymbol{p}_2) - 3\cos\left(\boldsymbol{p}_1, \widehat{\boldsymbol{R}}\right)\cos\left(\boldsymbol{p}_2, \widehat{\boldsymbol{R}}\right)\right], \end{aligned} \tag{1.80}$$

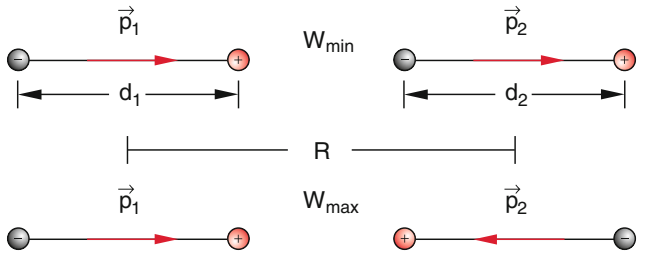


Fig. 1.55 Interaction energy of two dipoles for parallel and anti-parallel orientations

which allows one to calculate the force between the dipoles using $\boldsymbol{F} = -\mathbf{grad}\, W_{\text{pot}}$. This illustrates that the interaction between two dipoles decreases with the third power ($\propto 1/R^3$) of the distance between the midpoints of the dipoles and it depends furthermore on the mutual orientation of the two dipoles.

The same result can be derived from $\boldsymbol{F} = \boldsymbol{p}_1 \cdot \text{grad}\, \boldsymbol{E}_2$ in Eq. (1.29) where E_2 is the electric field generated by p_2 which is outlined in Eqs. (1.25a, 1.25b).

The dipole-dipole interaction has a minimum

$$W_{\text{min}} = -\frac{2p_1 p_2}{4\pi\varepsilon_0 R^3}$$

for the collinear orientation and a maximum

$$W_{\text{max}} = \frac{2p_1 p_2}{4\pi\varepsilon_0 R^3}$$

for anti-collinear arrangements (Fig. 1.55). This implies that two suitable oriented dipoles attract each other. Figure 1.56 shows the general case and five special cases of dipoles with different mutual orientations.

In gases or liquids the orientation of the molecular dipoles are randomly distributed because of thermal motions of the molecules. Therefore the macroscopic dipole moment per unit volume of all N molecules at room temperature is zero (Fig. 1.57).

In an external field $\boldsymbol{E}$ a torque acts upon the individual dipole molecules which is proportional to $|\boldsymbol{E}|$ and turns the molecules with increasing field more and more into the direction of minimum total energy..

The macroscopic polarization $\boldsymbol{P} = \frac{1}{V}\sum \boldsymbol{p}_i$ of polar but randomly oriented molecules increases proportional to E until all molecules are oriented in the same direction.

For a given field E the orientation of the moments becomes better at low temperatures A measure of the orientation is the ratio

$$v = \frac{pE}{3kT}$$

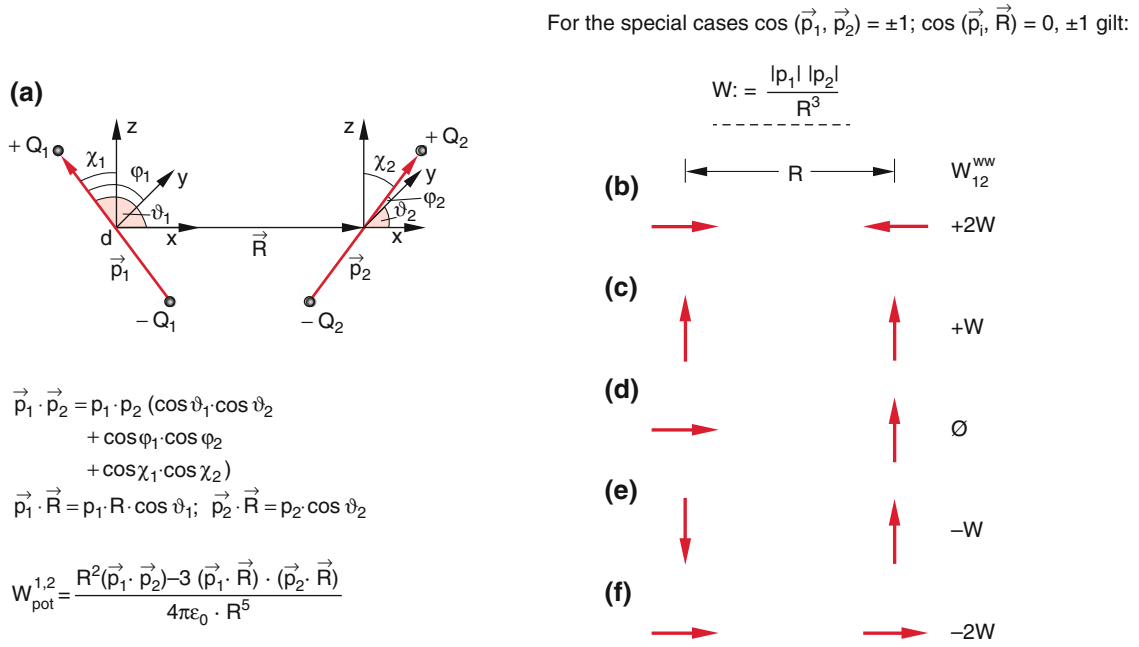


Fig. 1.56 Two interacting dipoles **a**) with arbitrary mutual orientation, **b**–**f**) for some special orientations

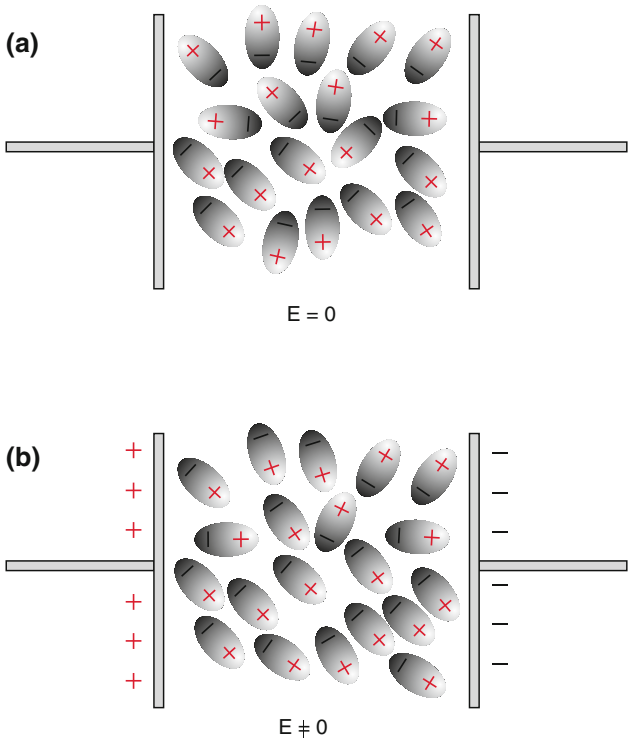


Fig. 1.57 Random orientation of dipoles at temperature T **a**) without external field **b**) partial orientation with external field

of orienting electrostatic energy and the disorienting thermal energy 3kT. Considering the statistical mean at the temperature T only a fraction $v<1$ of all molecules is oriented parallel to the field direction. A more rigorous consideration gives the same result for the macroscopic polarization (see Vol. 3)

$$\boldsymbol{P} = \frac{Np^2}{3kT}\boldsymbol{E}. \tag{1.81}$$

In an external field also polar molecules get an additional induced dipole moment that is proportional to $\boldsymbol{E}$ so that the total polarization becomes

$$\boldsymbol{P} = (a + b \cdot |\boldsymbol{E}|)\boldsymbol{E} \tag{1.81a}$$

In technical realized fields $\boldsymbol{E}$ is generally $b \cdot |\boldsymbol{E}| \ll a$.

Example

The electric dipole moment of water molecules H_2O is $p = 6.1 \times 10^{-30}$ C m. The attracting interacting force between two water molecules has a maximum if both dipole moments are parallel to the line of connection between the two dipole centers.

At a distance of $R = 3 \times 10^{-10}$m we obtain from (1.80)

$$W_{\text{pot}} = -2.3 \times 10^{-20}\text{J} = 140\,\text{meV}.$$

Since the thermal motions of the molecules causes a random orientation of the molecules mean interaction energy becomes smaller by about a factor of ten.

Many molecules (e.g. H_2, CO_2) have no permanent dipole moment i.e. the factor a in (1.81a) is zero. Here the total polarization of gases and liquids increases proportional to E^2.

The interaction between polar molecules (dipole-dipole interaction) and between non-polar molecules (induced dipole-induced dipole interaction) plays an important role in molecular physics because it contributes an important part to the chemical bond.

Summarizing we can state

The reasons for the macroscopic polarization of matter in an electric field are:

(a) In an electric field the charges of molecules are displaced and therefore the molecules get an induced dipole moment.
(b) The spatial orientation of polar molecules with permanent dipole moments which are randomly oriented without external field, are oriented with increasing external field into a preferential direction.

1.9 Electrostatics in Nature and Technology

Phenomena of electrostatics play an important role in a lot of fields of our surrounding nature as well as to solve technical problems. This we will illustrate by a few examples.

1.9.1 Triboelectricity and Contact Potential

When two different bodies ar brought into close contact e.g. by rubbing one against the other, electrons are transferred from one body to the other and after the bodies are separated they bear opposed charges (triboelectrics, Fig. 1.58). The direction of transfer of charges is determined by the difference of the effective work function of the electrons in the specific material. The electrons are transferred from the body with the lower work function to that of higher work function because then the electrons gain energy.

This separation of charges causes a potential difference $U = \Delta\phi$ between the bodies which is called ***contact potential***. The different materials can be ordered according to their difference of contact potentials between a reference material and the material considered. Table 1.3 lists the materials in the order of increasing work functions (electrochemical series). After the separation of two contacting materials that with the lower work function has the positive charge.

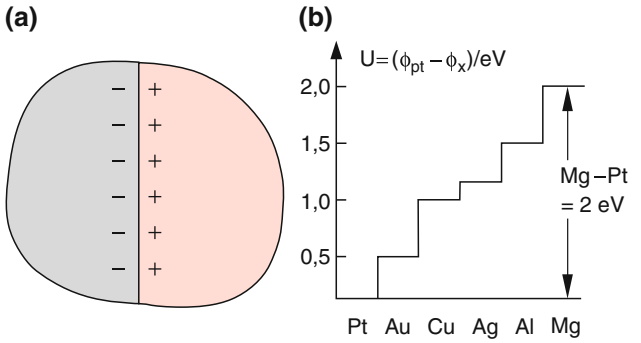


Fig. 1.58 **a**) Origin of tribo-electricity **b**) electric work functions of some metals

Table 1.3 Work functions of some elements, ordered according to the electrochemical series. (Stöcker: Taschenbuch der Physik, Harri Deutsch, Frankfurt)

Metal	ϕ/volt	Metal	ϕ
Cs	2.14	Pb	4.25
Rb	2.16	Al	4.28
K	2.30	Sn	4.31
Sr	2.59	Zn	4.33
Ba	2.70	Ag	4.52
Na	2.75	W	4.55
Ca	2.87	Mo	4.6
Li	2.90	Fe	4.63
Nd	3.30	Cu	4.65
Th	3.47	Au	5.1
Mg	3.66	Ni	5.15
Ti	3.87	Pd	5.40
Cd	4.22	Pt	5.66

Note: For technical applications triboelectricity often has a negative image because it can play a dangerous role. For example during the flow of liquids (oil, gasoline) into ships or when grain is blown into granaries explosions can occur. To suppress these explosions charging has to be avoided.

1.9.2 The Electric Field of Our Atmosphere

Even at fair weather our earth generates an electric field in the atmosphere. It is directed towards the surface of the earth and decreases rapidly, faster than $1/h^2$ with increasing height h above ground (Fig. 1.59).

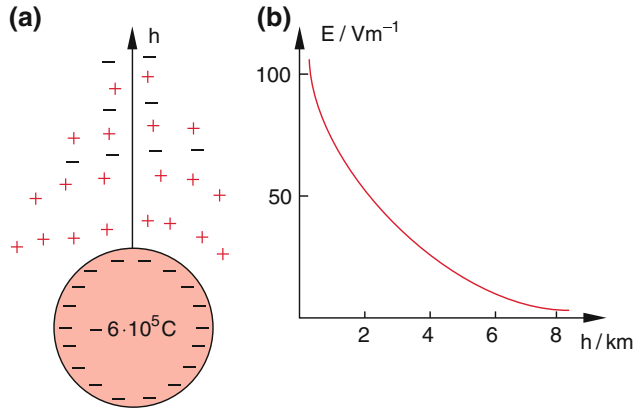


Fig. 1.59 **a**) Charge distribution of the earth surface and in the atmosphere **b**) Electric field as a function of the heights h above ground

Quantitative measurements show that the field strength a few meters above the surface has a mean value of $E = 130\,\mathrm{V/m}$ [6]. From these measurements we conclude.

- The earth has a temporally averaged negative charge of about $Q = -6 \times 10^5\,\mathrm{C}$.
- The atmosphere contains carriers of positive as well as negative charges with a surplus of positive charges in the lower layers which partly screen the electric field of the earth and cause the rapid decrease of the electric field E with height h.
- The electric field of the earth accelerates the carriers of positive charges toward the surface. This causes a current density of about $2 \times 10^{-12}\mathrm{A/m^2}$ and a total current of about 10^3 A toward the whole earth and thus the surplus of negative charges is reduced. If this were the only mode of charge transportation the total surplus of charges of the earth would vanish within about 10 min.
- The mean value of the electric field of the earth is constant over long times and therefore the charge of the earth must be also constant. The inflow of positive charges must be compensated by a corresponding inflow of negative charges resp. by a drain of positive charges [7].

This can happen by a vertical wind that transports positively charged dust particles above land or water droplets above sea into the higher atmosphere. Also lightings cause a balance of charges between clouds and the surface of the earth.

- The density of ions (positive and negative ions) in the atmosphere depends strongly on weather. During fine weather the average number of pairs of ions is 10^6–10^8 $\mathrm{m^{-3}}$ compared to the density of the neutral molecules of $10^{25}\mathrm{m^3}$. Thus the ionization of the atmosphere is only very weak. This situation changes in the ionosphere ($h > 70\,\mathrm{km}$) where UV-radiation and particle radiation of the sun initiate photoionization and a large part of the gas molecules become ionized.

1.9.3 The Generation of Lightnings

Lightnings are caused if hot and cold air currents meet and very strong vertical currents of air are created. This transport of charged particles of dust, ice and especially of water droplets creates local differences of charge densities. High electric fields are the result. These vertical currents of wet air between regions of large differences of temperature lead to condensation of water if it is transported from hot to cold areas resp. to vaporization for the opposite direction. The impressive result are great cumulus clouds. The upper layers of the cloud are positively charged while the lower layers bear more negative charges. This is caused by the electric field of the earth that induces a dipole moment in water droplets with the positive charges facing the earth. The greater drops of water sink because of their mass and become mostly negatively charged because an impact with positive ions is more likely at the downward face of the drops than for its backside (Fig. 1.60). Smaller drops move with the rising air and become mostly positively charged (for the same reason).

If this separation of charges has lead to a sufficiently high electric field strength between the upper and lower layers of a cloud or between cloud and surface of the earth an electric discharge (flash of lightning) occurs. The average amount of charges transported in a flash is about 10 C often within 10^{-4} s (10^5 A!!) and results in a compensation of the charge differences. This high electric current heats the air locally, which results in a pressure shock wave (thunder). Between the upper layer of a lightning cloud and the ionosphere in a height between 50–100 km voltages up to 40 MV can arise. In such high fields electrons can be accelerated to such high

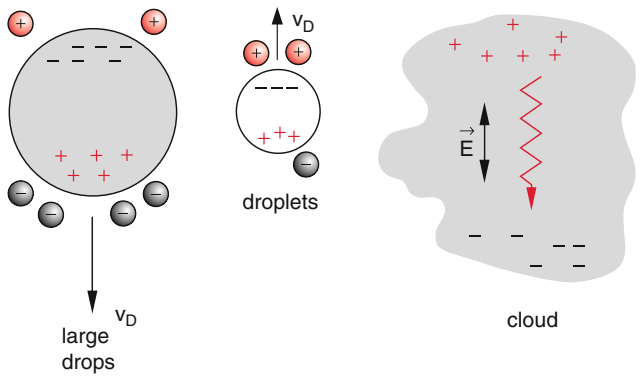


Fig. 1.60 Generation of lightenings by charge separation in water droplets

energies, that for collisions of these electrons with atoms or molecules X-rays or even gamma radiation is generated. ($h \cdot \nu \approx 30\,\text{MeV}$) [8, 9].

The energy of lightning is $W = I \cdot U \cdot \Delta t = 10^5 \times 4 \times 10^7 \times 10^{-4}\,\text{Ws} = 4 \times 10^8\,\text{Ws} = 400\,\text{MWs}$.

After the lightning has reduced the voltage difference it takes only a few seconds to rebuild the voltage differences again by transport of charge carriers through air currents.

1.9.4 Ball Lightnings

A great number of persons among them also recognized scientists report of observations of bright shining gas balls of diameters between a few centimeters and nearly one meter. These balls exist for several seconds, moving through the air, cause serious burns or electric shocks at contact like usual flashes of lightnings. Therefore they were called ball lightnings. Because up to now there exists no scientific explanation for this phenomenon it has been neglected or even consigned to the realms of fantasy. During the last years scientists again have taken care of the problem and tried to investigate ball lightnings under controlled physical conditions.

With restrictions several laboratories have been successful in creating ball lightings by discharging a high voltage capacitor into water containing salt. The experimental arrangement is shown in Fig. 1.61. A ring-shaped and a U-formed electrode are immersed into salt water. A Capacitor (1 mF) is charged up to 5 kV and then discharged through the salt water by opening a switch. The U-shaped electrode is isolated except at its upper end which is located at the surface of the salt water. Here a bright light phenomenon appears shaped as a very brilliant sphere with about 20 cm diameter and a temperature of about 10.000 K. The bright ball exists for about 0.3 s. It is named by the observers as plasmoid, because it contains electrons, ions and water molecules. The results show that ball lightnings are quasi neutral plasma spheres of electrons and positive ions under metastable conditions, which can exist for a while before they explode by recombination [10]. Until now it cannot be explained why the hot ball exists for a longer time as expected. For more information about ball lightnings more detailed experiment are necessary [11, 12].

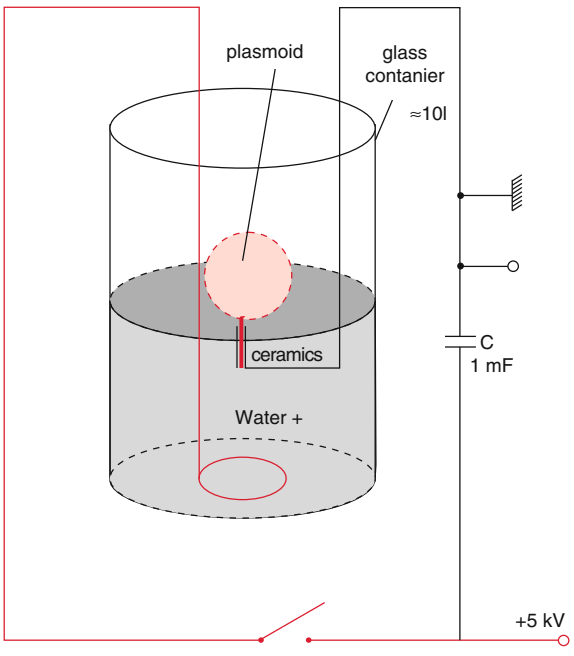


Fig. 1.61 Experimental arrangement for the greation of ball ligthenings

1.9.5 Electrostatic Air Filter

The emission of dust produced by power stations and industrial plants can be reduced significantly by electrostatic air filters. Figure 1.62 shows schematically a possible version. Inside the exhaust channel a central wire is mounted between that and the surrounding metallic tube a high voltage (50–100 kV) is applied. The resulting electric field ignites a gas discharge. By adsorption of mostly negative charges on the dust particles these are deposited at the positively charged plates from where they are removed mechanically and collected at the bottom [13, 14].

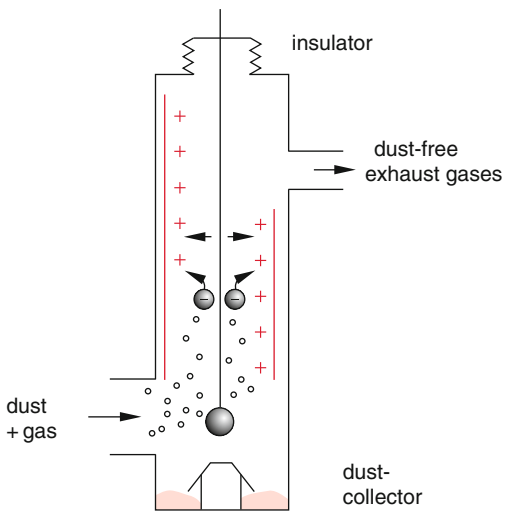


Fig. 1.62 Electrostatic dust filter

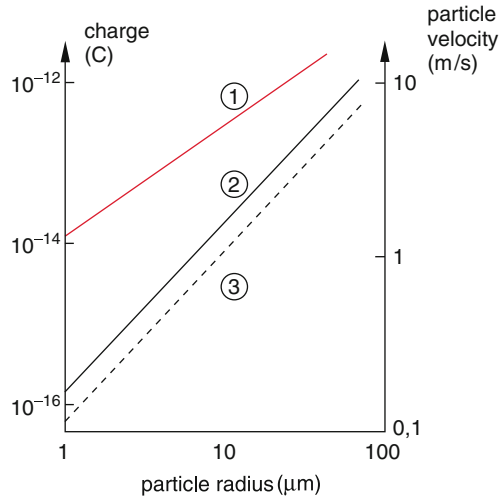


Fig. 1.63 Maximum achievable charge of a droplet in a coronary discharge as a function of the droplet radius (left scale). Velocity of droplets in an electric field $E = 5 \times 10^5$ V/m (right scale) (1) water (2) conductive spheres (3) dielectric spheres (ref: [16–18])

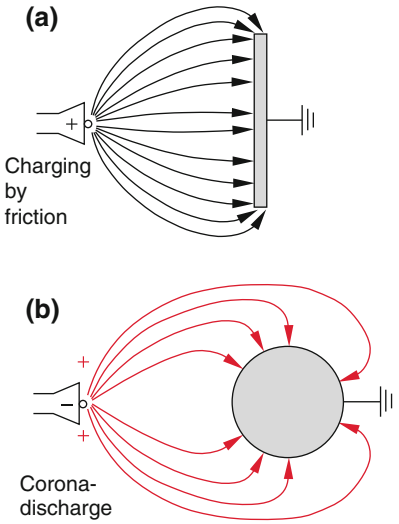


Fig. 1.64 **a**) Dye-coating of a plane surface by dye droplets ejected from a nozzle and charged by friction **b**) coating of a spherical surface

1.9.6 Electrostatic Deposition of Dye Coating

Through a nozzle a solution of colored dye is sprayed to a conducting surface. Depending on the type of dye the drops are charged by rubbing at each other or by a coronary discharge. Figure 1.63 shows the maximum charge per drop as a function of the drop radius consisting of water, conducting particles, or insulators. An electric field accelerates the particles until their drag in air is compensated.

In this case the Stokes's friction force (Vol. 1, Chap. 8) and the accelerating force in an electric field are equal but opposite.

$$F = 6\pi \cdot \eta \cdot r \cdot v = q \cdot E$$

The stationary speed v is then

$$v = \frac{q}{6\pi\eta r} E.$$

The object to be coated is grounded in order to deposit the particles that move along the field lines at the object (Fig. 1.64). By an appropriate form of the electrodes the spatial distribution of the field can be adapted and thus the distribution of dye particles be optimized. In each step of the deposition only one color is used. In successive steps a multi-color coating can be realized [15].

1.9.7 Electrostatic Copier and Printer

The principle of electrostatic copiers has been invented already in 1935. It is based on a combination of the photo-electric properties of certain materials (selenium, zinc-oxide) and the electrostatic deposition of a fine powder on charged surfaces.

The process of copying or printing is shown in Fig. 1.65. A cylindrical drum coated with electrically charged selenium is kept in the dark. An optical system creates an image of the object to be copied on the drum. A part of the charges is removed due to this exposure. The number of the removed electrons is proportional to the incident light intensity (photo-effect). The surface charge on the drum is larger at those areas that correspond to the dark areas of the object than at the bright areas. Powder with opposite charge is accelerated onto the drum and remains at the charged positions.

A charged sheet of paper is pressed against the rotating drum where it removes the charged powder from the drum. Then the paper runs through a heater where the melted powder is permanently burned into the paper. The drum then passes a blade and a brush that removes the remaining powder to have a clean surface for the next copy [16, 17].

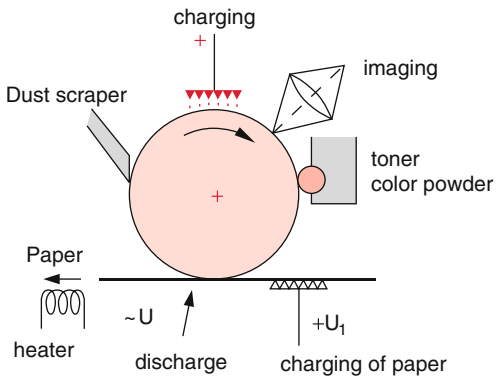


Fig. 1.65 Principle of a photo-copier

Today selenium has been replaced by organic semiconductors, e.g. polyvinyl-carbazol, as coating of the drum. These materials are made conductive by doping. They are cheaper and not poisonous as selenium and also the quality of the reproduction is higher.

1.9.8 Electrostatic Charging and Neutralization

Newspapers and piles of newspapers are charged to avoid slipping during their transport. Address labels are charged to tie them to newspapers or leaflets before they are packed into plastic wrap. When a surface is to be covered very evenly by a liquid, a previous charging is very useful. An example for this application is the production of DVDs, which consist of two thin plastic disks that are glued together.

On the other hand for a lot of applications it is important to protect the used materials from possible charging. One example is the discharging of foils that are charged unintentionally during their production. To use them as packaging they mostly have to be electrically neutral. Therefore the foils are discharged before they are wound up on rolls. The principle is shown in Fig. 1.66. Electrons are sprayed onto the foil and a grounded roller removes the excess charge.

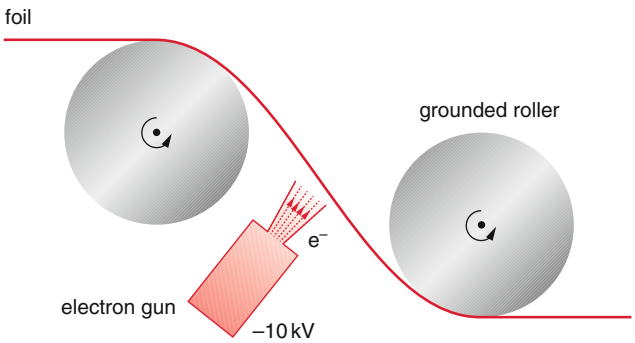


Fig. 1.66 Neutralization of charged foils

Summary

- The static electric field is created by charges. With N charges Q_i at the positions $\boldsymbol{r}_i$ the total electric field strength at the observation point $P(\boldsymbol{R})$ is

$$\boldsymbol{E}(P) = \frac{1}{4\pi\varepsilon_0}\sum_{i=1}^{N}\frac{Q_i}{|\boldsymbol{R}-\boldsymbol{r}_i|^3}(\boldsymbol{R}-\boldsymbol{r}_i).$$

- The field of a spatially distributed charge density $\varrho(\boldsymbol{r})$ at the point $P(\boldsymbol{R})$ outside the volume of the charges is

$$\boldsymbol{E}(P) = \frac{1}{4\pi\varepsilon_0}\int\frac{\varrho(\boldsymbol{r})(\boldsymbol{R}-\boldsymbol{r})}{|\boldsymbol{R}-\boldsymbol{r}|^3}\mathrm{d}^3r.$$

- The electrostatic field is conservative and therefore can be written as the gradient

$$\boldsymbol{E} = -\mathbf{grad}\,\phi$$

of the scalar potential

$$\phi(\boldsymbol{R}) = \frac{1}{4\pi\varepsilon_0}\int\frac{\varrho(\boldsymbol{r})\,\mathrm{d}^3r}{|\boldsymbol{R}-\boldsymbol{r}|}$$

- The charges are the sources of the electric field. In vacuum the Poisson equation is valid

$$\mathrm{div}\,\boldsymbol{E} = \varrho/\varepsilon_0 \Rightarrow \Delta\phi = -\varrho/\varepsilon_0,$$

where Δ is the Laplace operator.

- Inside of dielectric matter the electric field strength decreases. For the field strength $\boldsymbol{E}$ and the displacement field $\boldsymbol{D}$ is

$$\begin{aligned}\boldsymbol{E}_{\mathrm{Diel}} &= \frac{1}{\varepsilon}\boldsymbol{E}_{\mathrm{Vac}},\\ \boldsymbol{D}_{\mathrm{Diel}} &= \varepsilon_0\boldsymbol{E}_{\mathrm{Vac}} = \varepsilon\varepsilon_0\boldsymbol{E}_{\mathrm{Diel}},\\ \mathrm{div}\,\boldsymbol{D} &= \varrho.\end{aligned}$$

- At a boundary between two media with their relative dielectric constants ε_i is

$$\begin{aligned}\boldsymbol{E}_{\parallel}^{(1)} &= \boldsymbol{E}_{\parallel}^{(2)}, \quad \boldsymbol{D}_{\perp}^{(1)} = \boldsymbol{D}_{\perp}^{(2)}\\ \frac{1}{\varepsilon_1}\boldsymbol{D}_{\parallel}^{(1)} &= \frac{1}{\varepsilon_2}\boldsymbol{D}_{\parallel}^{(2)}, \quad \varepsilon_1\boldsymbol{E}_{\perp}^{(1)} = \varepsilon_2\boldsymbol{E}_{\perp}^{(2)}.\end{aligned}$$

- The electric flux through the surface S

$$\Phi_{\mathrm{el}} = \oint\boldsymbol{E}\cdot\mathrm{d}\boldsymbol{S} = Q/\varepsilon_0$$

is a measure for the source strength of the charge Q enclosed by the surface S.

- From $\boldsymbol{E} = -\nabla\phi$ follows $\mathbf{rot}\,\boldsymbol{E} = \boldsymbol{0}$.

The static electric field is non-rotational it is eddy-free. There are no closed field lines. Field lines begin on positive charges and end on negative ones.

- Two conducting surfaces represent a capacitor with a capacitance C that can carry a charge $Q = CU$. The capacitance of a parallel plate capacitor with plate area A and plate separation d is

$$C = \varepsilon\varepsilon_0\cdot\frac{A}{d}.$$

- The capacitance of a spherical capacitor of radius R is

$$C = 4\pi\varepsilon_0 R.$$

- The force $\boldsymbol{F}$ acting on a test charge q in an electric field E is

$$\boldsymbol{F} = q\cdot\boldsymbol{E}.$$

- In an electric field $\boldsymbol{E}$ the work W that has to be performed to bring the charge q from the position P_1 to the position P_2 is

$$\begin{aligned}W &= q\int_{P_1}^{P_2}\boldsymbol{E}\cdot\mathrm{d}\boldsymbol{s}\\ &= q(\phi(P_1)-\phi(P_2)) = q\cdot U,\end{aligned}$$

where the voltage U equals the potential difference $\Delta\phi = \phi_1 - \phi_2$.

- An electric dipole consists of two charges of opposite signs, $-Q$ and $+Q$ at a distance d. Its dipole moment is

$$\boldsymbol{p} = Q\cdot\boldsymbol{d},$$

where $\boldsymbol{d}$ points from the negative to the positive charge. In a homogeneous field a torque acts

$$\boldsymbol{D} = (\boldsymbol{p}\times\boldsymbol{E}).$$

- In an inhomogeneous field an additional force acts

$$\boldsymbol{F} = \boldsymbol{p}\cdot\mathbf{grad}\,\boldsymbol{E}.$$

- The potential $\phi(P)$ and the field $\boldsymbol{E}(P)$ of an arbitrary charge distribution that is sufficiently far away from the point P can be calculated by a series expansion (multipole expansion).
- Also between charge distributions that are neutral to the outside, forces act except the distributions are spherically symmetric.

- An electric field displaces charges in matter. This displacement is called influence for conductors and polarization for insulators. There exists no field inside a conductor. Inside insulators the field decreases to $E_{\mathrm{Diel}} = \frac{1}{\varepsilon} E_{\mathrm{Vac}}$ because induced dipoles are generated and their field reduces the external field.
- The dielectric polarization

$$\boldsymbol{P} = N \cdot q \cdot \boldsymbol{d} = N \cdot \alpha \cdot \boldsymbol{E}_{\mathrm{Diel}}$$

is equal to the sum of all induced dipoles per unit volume and it is also proportional to the field strength E_{Diel}. The factor α depends on the material and is called polarizability.

- The static electric field in matter or in vacuum is completely described by the field equations

$$\mathbf{rot}\, \boldsymbol{E} = \mathbf{0}, \quad \mathrm{div}\, \boldsymbol{D} = \varrho, \quad \boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}.$$

Problems

1.1 Gedankenexperiment: Two small balls of sodium of mass $m_1 = m_2 = 1\,\text{g}$ are separated by 1 m. Suppose that for every tenth atom one electron is missing. Which surface charge density σ is present on each ball and with which force the balls repel each other? (density of sodium $\varrho = 0.97\,\text{g/cm}^3$, mass of a sodium atom $m_s = 23 \times 1.67 \times 10^{-27}\,\text{kg}$, elementary charge $e = -1.602 \times 10^{-19}\,\text{C}$).

1.2 Two equal balls of mass m carry equal charges Q and are suspended by strings with length L with the same suspension point A (Fig. 1.67).
(a) How large is the angle φ? (numerical example: $m = 0.01\,\text{kg}$, $l = 1\,\text{m}$, $Q = 10^{-8}\,\text{C}$.)
(b) How large is φ, if a conducting plate with the surface charge density $\sigma = 1.5 \times 10^{-5}\,\text{C/m}^2$ is placed symmetrically between the two balls?

1.3 A conducting annulus disc with the inner radius R_1 and the outer radius R_2 carries the charge density σ.
(a) Calculate the force on a point charge q which is located at the distance x on the symmetry axis perpendicular to the disc
(b) What is the result for the limiting cases (α) $R_1 \to 0$, (ß) $R_2 \to \infty$, (γ) $R_1 \to 0$ and $R_2 \to \infty$?

1.4 Two conducting balls with radius R_1 and R_2 are connected by a thin conducting wire with length $L \gg R_1, R_2$. When the charge Q is brought onto the system, how is the charge distribution Q_1 and $Q_2 (Q_1 + Q_2 = Q)$ on the two balls and what are the electric fields E_1 and E_2 on the two surfaces?

1.5 A point charge Q_1 is located at the point $P_1(0, 0, z = a)$ another charge Q_2 at the point $P_2(0, 0, z = -a)$. Calculate the force onto a charge q at the point $P(r, \vartheta, \varphi)$ and the potential energy E_{pot} for the two numerical examples $Q_1 = Q_2 = 10^{-9}\,\text{C}$ and $Q_1 = -Q_2$. What are the first three members of the multipole expansion?

1.6 Calculate the potential energy of the three charge distributions shown in Fig. 1.68, i.e. the energy, one has to spend in order to bring the three charges from infinity to the configurations shown in Fig. 1.68?

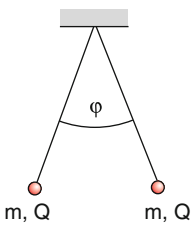


Fig. 1.67 Illustration of problem 1.2

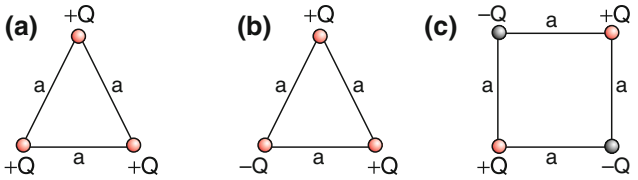


Fig. 1.68 Illustration to solution 1.6

1.7 Calculate the quadrupole moments of the charge distributions in Fig. 1.69.

1.8 Calculate the potential $\phi(r)$ and the electric field $\boldsymbol{E}(r)$ for a charged non conducting homogeneous ball (radius R, charge Q). What is the work necessary to bring a charge q
(a) from $r = 0$ to $r = R$
(b) from $r = R$ to $r = \infty$ if we choose $\phi(r = \infty) = 0$.

1.9 Execute the differentiation in (1.34) leading to the multipole expansion, in all details.

1.10 Show that for a charged homogeneous ball with the total charge Q all terms in (1.35) are zero, except the monopole term.

1.11 For high voltage lines 4 wires (each with radius R) are used that are arranged in such a way, that their intersections with a plane $z = z_0$ mark 4 points P_i at the locations $(x = \pm a, y = 0)$; and $x = 0, y = \pm a)$ which form a square with side lengths $a \cdot \sqrt{2}$. All wires have the same voltage U against earth. Calculate
(a) the electric field on the diagonal lines $x = 0$ and $y = 0$
(b) the electric field $\boldsymbol{E}(r, \varphi)$ on the surface of each wire with radius $R = a/8$
(c) by which factor is E smaller than for a single wire at the voltage U?
Numerical values: $R = 0.5\,\text{cm}$, $a = 4\,\text{cm}$, $U = 3 \times 10^5\,\text{V}$

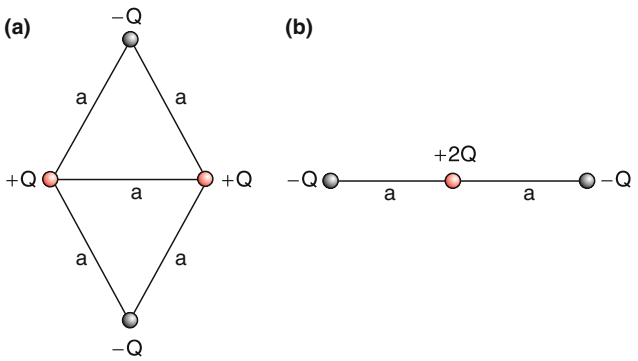


Fig. 1.69 Illustration to problem 1.7

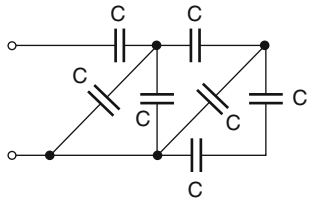


Fig. 1.70 Illustration to problem 1.13

1.12 Between the two plates of a plane capacitor (distance d = 1 cm; area $A = 0.1\ \mathrm{m}^2$) a voltage U = 5 kV is applied.
- (a) How large are capacity C, charge Q on the plates and electric field E?
- (b) Prove, that the field energy of the capacitor is $W = \frac{1}{2} C \cdot U^2$
- (c) What is the torque on an atomic dipole ($q = 1.6 \times 10^{-19}\ \mathrm{C}, d = 5 \times 10^{-11}\ \mathrm{m}$) in the electric field of a capacitor with the dipole axis parallel to the plates of the capacitor? How much energy is gained resp. lost, if the dipole axis is parallel or antiparallel to the electric field vector?

1.13 What is the total capacity of the design, shown in Fig. 1.70?

1.14 The charge Q is applied to the left plate of the capacitor circuit in Fig. 1.71. What are field-distribution $\boldsymbol{E}(x)$ and potential $\phi(x)$?

1.15 On both sides of the cylindrical capacitor with radii R_1 and R_2 and angular extension φ (Fig. 1.72) are apertures with a slit at $R = (R_1 + R_2)/2$.
- (a) Which voltage U has to be applied that allows an electron with velocity v_0 to pass both slits?
- (b) At which angle φ is the capacitor focusing, i.e. electrons with a small angle against the path R = const. should also pass the exit slit?

1.16 A thin wire with length L is bent to a circular arc with R = 0.5 m. It carries the charge Q. Determine amount and direction of the electric field in the center of curvature as a function of the ratio L/R.

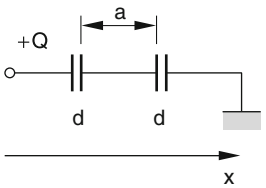


Fig. 1.71 Illustration to problem 1.14

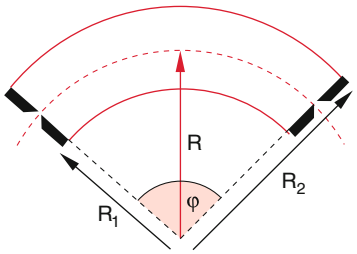


Fig. 1.72 Illustration to problem 1.15

References

1. A. B. Arons: Development of Concepts of Physics (IAddison Wesley Reading 1965)
   JED Z. Buchwald, Robert Fox: The Oxford Handbook of the History of Physics Oxford University Press (2013)
2. J. Munro: The story of electricity (Indy Publ. New York 2008)
3. Weltner/john/Weber: Mathematics for Physicists (Springer 2014) G. Arfken et.al. mathematical Methods for Physicists 7th ed. (Academic Press 2012)
4. Van de Graaff, R. J.; Compton, K. T; Van Atta, L. C. (February 1933). "The Electrostatic Production of High Voltage for Nuclear Investigations" (PDF). *Physical Review*. **43** (3): 149–157. https://en.wikipedia.org/wiki/Van_de_Graaff_generator
5. Millikan, R. A. "On the Elementary Electric charge and the Avogadro Constant". *Phys. Rev.* **2** (2): 109– 143 (1913). https://www.aps.org/programs/outreach/history/historicsites/millikan.cfm . https://en.wikipedia.org/wiki/Oil_drop_experiment
6. https://en.wikipedia.org/wiki/Atmospheric_electricity
7. F. K. Lutgens, E. J. Tarbuck: The Atmosphere 14th ed. (Pearson 2018)
8. H. Volland: Atmospheric Electrodynamics (Springer, Heidelberg 1984)
9. M. A. Uman: Ligthning (Dover Publ. 2003)
10. Cixing Liu, J. Martinsen: Ball Lightning (head of Zeus 2018). https://en.wikipedia.org/wiki/Ball_lightning
11. G. Fussmann: Künstlicher Kugelblitz: Phys. in uns. Zeit 39, issue 5 246 (2008)
12. http://amasci.com/tesla/ballgtn.html#res
13. https://en.wikipedia.org/wiki/Electrostatic_precipitator
14. K. R. parker: Applied electrostatic Precipator (Springer, Dordrecht 1997)
15. J.F. Hughes: Electrostatic Powder Coating (DEncyclopedia of Physical Sciences and Technology 2nd ed. Vol. 5 (Academic Press, New York 1992 p.839 ff)
16. https://en.wikipedia.org/wiki/Photocopier
17. https://www.xerox.com/downloads/usa/en/s/Storyofxerography.pdf
18. K.F. Riley et.al.: Mathematical Methods for Physicists (Cambridge University Press 2007)

Electric Currents

2

In this chapter we are going to introduce the basic features of stationary electric currents and their various actions as well as methods of their measurements. Especially the mechanisms of electric conductivity in solids, liquids and gaseous materials are explained. Also some possibilities for realizing electric current sources are represented.

2.1 Current as Transport of Charges

An electric current means the transport of electric charges through an electric conductor or in vacuum. The current I is defined as the amount of charges Q transported into one direction through a cross section of the conductor during the unit of time.

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t}. \tag{2.1}$$

The electric current I is measured in Coulomb per-second, or amperes (named after *Andre Marie Ampere* (1775–1836, Fig. 2.1a, b)) who first discovered that forces are acting between current carrying wires) (Figs. 2.2, 2.3 and 2.4).

$$[I] = \text{ampere} = \text{A}.$$

Its official definition [NIST] is (see Vol. 1, Sect. 1.6.8).

> The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length with negligible circular cross-section, and placed 1 m apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} N/m of length.

We define the current density $\boldsymbol{j}$ as the current through the unit cross-section area perpendicular to $\boldsymbol{j}$. The total current I through the area A is

$$I = \int_A \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A}. \tag{2.2}$$

Fig. 2.1 **a**) Andre Ampere. **b**) Definition of the current density j

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_2

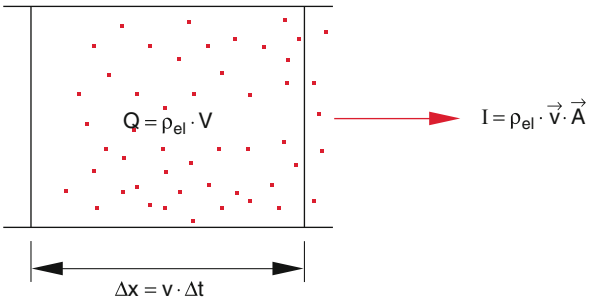


Fig. 2.2 Relation between electric current I and charge density ρ_{el}. The volume $V = \Delta x \cdot A$ where A is the cross section perpendicular to Δx

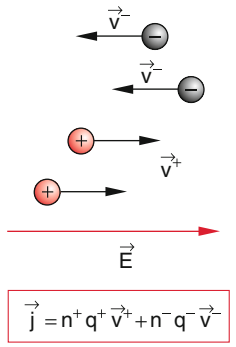


Fig. 2.3 Current density for conductors with charge carriers of both signs

If the current density is spatially constant then is $I = j \cdot \boldsymbol{A}$ (Fig. 2.1b).

The carriers of electric charges are mainly electrons and positive or negative ions that transport the charges. It depends on the material of the conductor which kind of charges predominates the transport. We distinguish:

- In metals mainly the free electrons contribute to the electric current.

Examples: solid and liquid metals, semiconductors

- In ionic conductors the carriers of charges are mainly ions.

Examples: electrolytes (acids, alkaline solutions, salt solutions), insulators with impurities.

- Mixed conductors, where electrons as well as ions contribute to the current.

Examples: gas discharges, plasma.

We consider a conductor with n charges q per unit volume that move with the velocity v in one direction. The charges inside the volume $V = \boldsymbol{A} \cdot \boldsymbol{v}\Delta t$ move during the time interval Δt through the cross-section area A of the conductor and represent the current

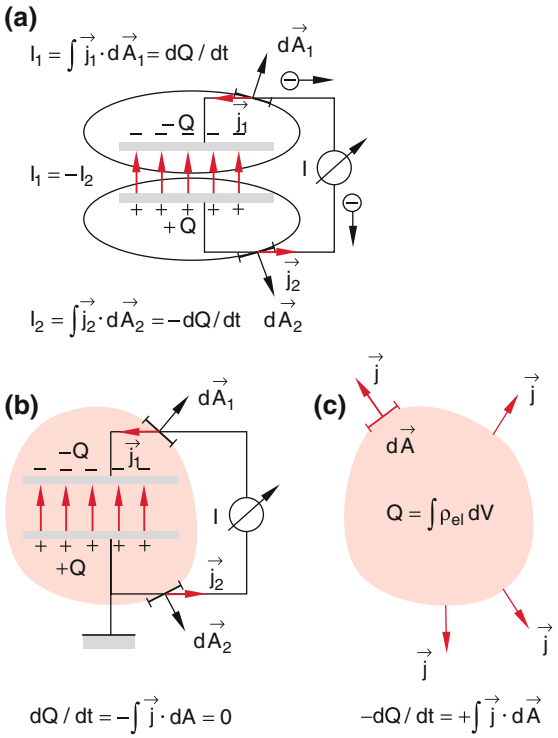


Fig. 2.4 Illustration of the continuity equation. **a**) Discharge of a capacitor where the electric current $I = \mathrm{d}Q/\mathrm{d}t$ flows through an arbitrary surface surrounding the capacitor **b**) illustration of continuity equation

$$I = n \cdot q \cdot \boldsymbol{A} \cdot \boldsymbol{v}$$

and the current density

$$\boldsymbol{j} = n \cdot q \cdot \boldsymbol{v}$$

With the charge density $\varrho_{el} = n \cdot q$ we can rewrite the current density as

$$\boldsymbol{j} = \varrho_{el} \cdot \boldsymbol{v} \tag{2.3}$$

If charges of both signs are present, e.g. in a gas discharge then the net charge density is

$$\varrho_{el} = \varrho_{el}^{+} + \varrho_{el}^{-} = n^{+}q^{+} + n^{-}q^{-},$$

and the total current density is

$$\boldsymbol{j} = n^{+}q^{+}\boldsymbol{v}^{+} + n^{-}q^{-}\boldsymbol{v}^{-}, \tag{2.3a}$$

In general the velocities $\boldsymbol{v}^{+}$ and $\boldsymbol{v}^{-}$ are of unequal amount and point into opposite directions. For example, in a gas discharge often is $q^{+} = e = -q^{-}$. Since the plasma is quasi-neutral it must be $n^{+} = n^{-} = n$ and the total current density becomes

$$\boldsymbol{j} = en(\boldsymbol{v}^{+} + \boldsymbol{v}^{-}) \quad |\boldsymbol{j}| = en(v^{+} - v^{-}) \tag{2.3b}$$

Note For historical reasons in electrical engineering the direction of current flow is defined as the flow direction of the positive charges although later on it has been discovered that the current in metals is caused by electrons (i.e. negative charge carriers). The "technical" direction of current is therefore from plus to minus.

The current through a closed surface $\boldsymbol{A}$ is

$$I = \oint \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A} = -\frac{\mathrm{d}Q}{\mathrm{d}t} = -\frac{\mathrm{d}}{\mathrm{d}t}\int \varrho_{\mathrm{el}}\mathrm{d}V \tag{2.4a}$$

and must be equal to the rate of decrease of the included charges.

With Gauss's theorem

$$\oint \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A} = \int \mathrm{div}\boldsymbol{j}\,\mathrm{d}V$$

we get the **continuity equation**

$$\mathrm{div}\boldsymbol{j}(\boldsymbol{r},t) = -\frac{\partial}{\partial t}\varrho_{\mathrm{el}}(\boldsymbol{r},t), \tag{2.4b}$$

This states that charges can neither be created nor destroyed.

The negative change rate of charges in the volume V is equal to the total flux through the surface of this volume.

2.2 Electric Resistance and Ohm's Law

In this section we will gain fundamental insight into the mechanism of charge transport in conductors and we will show the connection between electric field $\boldsymbol{E}$ and current density $\boldsymbol{j}$.

2.2.1 Drift Velocity and Current Density

Even without an external electric field $\boldsymbol{E}$ the free carriers of charges move inside a conductor. For example, the contribution of velocities of ions in a conducting liquid is determined by their thermal motions at the temperature T. Ions of mass m have a mean velocity (see Vol. 1, Chap. 7)

$$v = \langle|\boldsymbol{v}|\rangle = (8kT/\pi m)^{1/2}.$$

For example, Na-ions in a liquid have a mean velocity of 500 m/s without an external field and at room temperature T = 300 K.

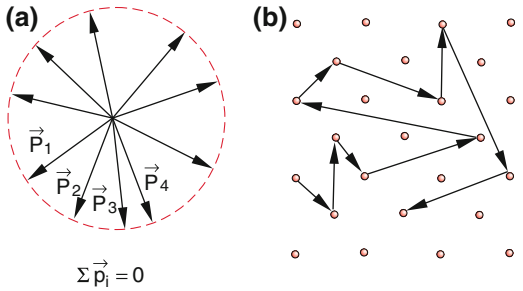


Fig. 2.5 **a** The average <p> of all electron momenta of the free electrons in the conduction band of metals is zero without external electric field. The peaks of the momentum vectors are located randomly on a sphere (Fermi-Sphere). The center of this sphere rests in momentum space. **b** Random path of an electron in an atomic lattice

Free electrons in metals have a much higher speed, about 10^6–10^7 m/s, caused by quantum mechanical effects.

On their way through the conductor the charge carriers collide very often with the atoms or molecules of the conductor. That changes the directions of the velocities statistically into all directions and without an external field the mean value $(\boldsymbol{v})$ becomes zero (Fig. 2.5). Therefore also the mean value of the current density is zero

$$\bar{\boldsymbol{j}} = n \cdot q \cdot \overline{\boldsymbol{v}} = \boldsymbol{0}.$$

The mean time interval τ between two successive collisions

$$\tau_{\mathrm{s}} = \Lambda/\bar{v}$$

is determined by the ratio of the mean free path Λ and the mean speed $\bar{v}$ (see Vol. 1, Chap. 7).

Example

1. Cu^{++} ions in a $CuSO_4$ solution at room temperature: $\bar{v} = 300$ m/s, mean free path $\Lambda = 10^{-10}$ m $\Rightarrow$ $\tau_{\mathrm{s}} = 3.3 \times 10^{-13}$ s.
2. Free electrons in Cu at room temperature: $\Lambda \approx 4 \times 10^{-8}$ m. Velocity at the Fermi surface $\bar{v} = 1.5 \times 10^6$ m/s $\Rightarrow$ $\tau_{\mathrm{s}} \approx 2.5 \times 10^{-14}$ s.

In an electric field $\boldsymbol{E}$ the carriers of charges with charge q and mass m suffer an additional force

$$\boldsymbol{F} = q \cdot \boldsymbol{E},$$

which leads to the acceleration $\boldsymbol{a} = \boldsymbol{F}/m$ (Fig. 2.6). During the time interval τ_{s} after the last collision the carriers get an additional speed

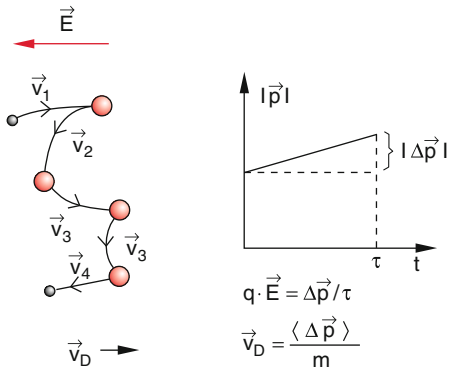


Fig. 2.6 Schematic representation of the electron path in a solid material, which is mainly determined by collisions with the atoms of the solid. An external electric field changes this path only slightly. The curvature of the path sections between two collisions is exaggerated. However, the mean momentum $\langle p \rangle \neq 0$

$$\Delta \boldsymbol{v} = a \cdot \tau_s = \frac{F}{m}\tau_s, \tag{2.5}$$

in the electric field which is in general small compared to its speed v and also small compared to the change $\Delta \boldsymbol{v}_i = \boldsymbol{v}_i - \boldsymbol{v}_{i-1}$ by the ith collision.

With the mean time τ_s after the last collision (= half mean time between two collisions). We get the mean additional speed $\langle \Delta v \rangle = (F/m)\langle \tau_s \rangle$. Without external field is $\langle \Delta \boldsymbol{v} \rangle = \boldsymbol{0}$.

This mean additional velocity

$$\boldsymbol{v}_D = \langle \Delta \boldsymbol{v} \rangle$$

is called **drift velocity**. Positive charges are transported into the direction of the field (negative charges opposite to the field) with a current density

$$\boldsymbol{j} = n \cdot q \cdot \boldsymbol{v}_D = \varrho_{el} \cdot \boldsymbol{v}_D. \tag{2.6a}$$

From (2.5) and (2.6a) together with $\boldsymbol{F} = q \cdot \boldsymbol{E}$ we get

$$\boldsymbol{j} = \frac{n \cdot q^2 \cdot \tau_s}{m}\boldsymbol{E} = \sigma_{el} \cdot \boldsymbol{E}. \tag{2.6b}$$

The quantity

$$\sigma_{el} = \frac{n \cdot q^2 \cdot \tau_s}{m} \quad \text{with} \quad [\sigma] = 1\,\mathrm{A\,V^{-1}\,m^{-1}}$$

depends on the material. It is called **electric conductivity**. It depends on the number density n of the charge carriers, on the mean time τ_s between two collisions, and on the mass m of the carriers. Often the drift velocity is written as

$$\boldsymbol{v}_D = u \cdot \mathbf{E} \quad \text{with} \quad u = \frac{\sigma_{el}}{n \cdot q}. \tag{2.6c}$$

where the factor $u = \sigma_{el}/n \cdot q$ is called **mobility** and has the unit of $m^2/(Vs)$. It represents the drift velocity in an electric field of $E = 1$ V/m.

Note In spite of the accelerating force $\boldsymbol{F} = q \cdot \boldsymbol{E}$ a constant drift velocity results. This is due to the fact, that the collisions consistently change the direction of the velocity. If $v_D \ll v$ all directions have the same probability immediately after the collision. The preference of the velocity direction into the field direction comes from the movement between two successive collisions i.e. in the time interval τ_s. During the collision the particle "forgets" its previous velocity direction.

The mean effect of the collisions can be described by a "friction force" $\boldsymbol{F}_F$ that is directed opposite to the field direction and compensates the field force $F_{el} = q \cdot E$ when the drift velocity is reached.

$$\boldsymbol{F}_f + \boldsymbol{F}_{EL} = \boldsymbol{F}_F + q \cdot \boldsymbol{E} = 0 \quad \langle \Delta \boldsymbol{v} \rangle = \boldsymbol{v}_D.$$

From (2.6b, 2.6c) we get

$$\boldsymbol{F}_f = -\frac{nq^2}{\sigma_{el}}\boldsymbol{v}_D = \frac{m}{\tau_s}\boldsymbol{v}_D.$$

The weaker the friction force $\boldsymbol{F}_f$ is, the higher becomes the conductivity σ_{el}.

In an electric field the current density $\boldsymbol{j} = \varrho_{el} \cdot \boldsymbol{v}_D$ is limited by the collisions of the charge carriers with the material of the conductor. The electric conductivity is determined by three facts:

(1) The concentration n of the charge carriers.
(2) The mean time interval τ_s between subsequent collisions.
(3) The mass m of the charge carriers.

Examples

1. In case of conduction by electrons in copper is $\sigma_{el} = 6 \times 10^7$ A/Vm, $n = 8.4 \times 10^{28}\,\mathrm{m^{-3}}$, and $q = -e = -1.6 \times 10^{-19}$ C. With these numbers the mobility becomes $|u| = 0.0043\,\frac{\mathrm{m/s}}{\mathrm{V/m}}$. For an electric field of 0.1 V/m a current of 600 A flows through 1 $\mathrm{cm^2}$ of copper wire. However, the electrons move only with a drift velocity of 0.4 mm/s!, whereas the mean speed of electrons in copper is $v = 1.6 \times 10^6$ m/s, which is about 0.5% of the speed of light. This illustrates that $v_D \ll \langle v \rangle$.
2. In an electrolytic conductor the mean thermal velocity of the ions is about 10^3 m/s. For a density of 10^{26} ions per $\mathrm{m^3}$ and a current density $j = 10^4$ A/$\mathrm{m^2}$ the drift velocity is

$$v_D = j/(n \cdot e) = 6 \times 10^{-4}\,\mathrm{m/s} = 0.6\,\mathrm{mm/s}$$

and therefore still very low compared to $\langle|\boldsymbol{v}|\rangle$. For a mobility $u = 6 \times 10^{-8}\,\mathrm{m^2/(Vs)}$ the electric conductivity is $\sigma_{el} = u \cdot n \cdot q \approx 1\,\mathrm{A/(Vm)}$. This is about eight orders of magnitude smaller then in copper.

It turns out that the electric conductivity of metals is proportional to the heat conductivity λ_h

$$\frac{\lambda_w}{\sigma_{el}} = a \cdot T \quad \textbf{Wiedemann–Franz' law}$$

The proportional constant $a \approx 3(k/e)^2$ is determined by the Boltzmann constant k and the elementary charge e. This shows that the free electrons in metals contribute both to the electric conduction as well as to the heat conduction (see Vol. 3).

2.2.2 Ohm's Law

Equation (2.6b) establishes the relation between current density **j** and electric field **E**. It is called
Ohm's law

$$\boldsymbol{j} = \sigma_{el} \cdot \boldsymbol{E}.$$

named after Georg Simon Ohm (1789–1854 Fig. 2.7). In a homogeneous conductor of cross-section area A and length L the integral form of Ohm' law can be written with $I = \int \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A} = j \cdot A$ and $U = \int E \cdot \mathrm{d}L$ as

$$I = \frac{\sigma_{el} A}{L} \cdot U = U/R. \tag{2.6d}$$

The **electric resistance** ***R*** of the conductor depends on the electric conductivity σ_{el} and the geometry of the conductor.

Fig. 2.7 Georg Simon Ohm

Table 2.1 Specific resistances ρ_s of some conductors and isolators

Material	$\varrho_s/10^{-6}\,\Omega\mathrm{m}$	Material	$\varrho_s/\Omega\mathrm{m}$
Silver	0.016	Graphite	1.4×10^{-5}
Copper	0.017	Water with 10 % H_2SO_4	2.5×10^{2}
Gold	0.027	H_2O + 10 % NaCl	8×10^{2}
Zinc	0.059	Teflon	1×10^{17}
Iron	≈0.1	Silicate glas	5×10^{15}
Lead	0.21	Porcelain	3×10^{16}
Mercury	0.96	Hard rubber	$\approx 10^{20}$
Brass	≈0.08		

$$R = \frac{L}{\sigma_{el} \cdot A} = \varrho_s \cdot \frac{L}{A} \quad \text{with} \quad \varrho_s = \frac{1}{\sigma_{el}} \tag{2.7}$$

The specific resistance $\varrho_S = 1/\sigma_{el}$ depends solely on the material of the conductor. The unit of the resistance R is

$$[R] = \left[\frac{U}{I}\right] = \frac{1\,\mathrm{Volt}}{1\,\mathrm{Ampere}} = 1\,\mathrm{Ohm} = 1\,\Omega.$$

The specific resistance $\varrho_s = R \cdot A/L$, with the unit $[\varrho_s] = 1\,\Omega\,\mathrm{m}$ is the resistance of a cube with $V = 1\ \mathrm{m}^3$. For practical reasons often the specific resistance is given as the resistance of a conductor with cross-section area 1 mm^2 and length 1 m. The unit is then $[\Omega\ \mathrm{mm^2/m}] = 10^{-6}\ \Omega$ m. Table 2.1 lists the specific resistance of a few materials.

If the specific resistance ϱ_s of a conductor is independent of I and U (Ohm's law) then current I and voltage $U = R \cdot I$ along the conductor are proportional, because R = const.

Note Along a conductor carrying the current I a voltage gradient appears (Fig. 2.8)

$$U(x) = \phi_1 - \phi(x) = R \cdot I \cdot \frac{x}{L} \tag{2.8}$$

- The conductor is no longer at a constant potential as in electrostatics and therefore its surface is no longer an equipotential surface.
- Not every conductor obeys Ohm's law. There are a lot of conductors where the conductivity σ_{el} depends on the current I and therefore in such cases the current is not proportional to the applied voltage (see Sect. 2.6).
- The electrical resistance R is also defined for conductors of complicated geometry where the ratio $R = U/I$ of voltage U across the connecting electrodes and the total current I defines the total resistance of the conductor. In such cases R cannot be readily

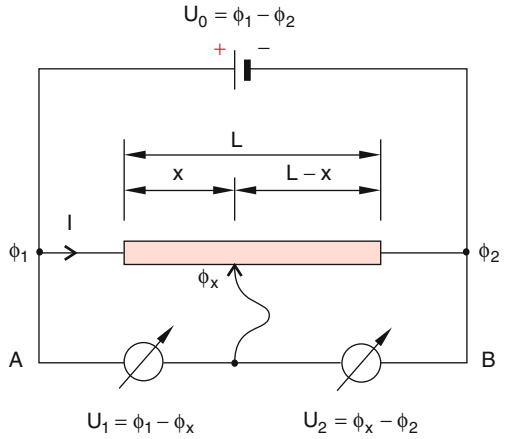


Fig. 2.8 Variable voltage divider

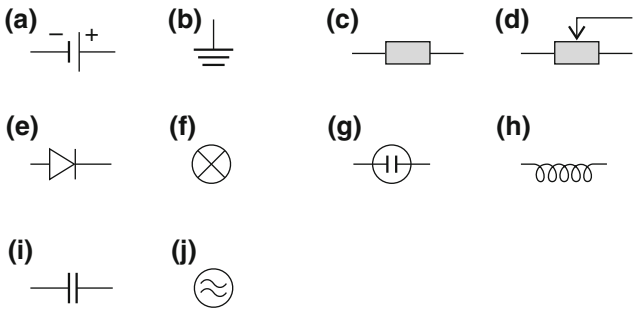


Fig. 2.9 Symbols for electrotechnical quantities: **a**) dc-voltage source **b**) ground **c**) resistor **d**) variable resistor **e**) rectifying diode **f**) light bulb **g**) voltage source **h**) inductance **i**) capacitor **j**) ac-source

calculated from the specific resistance ϱ_s and the complicated geometry of a sample but its value must be obtained from measurements.

- In Fig. 2.9 some symbols for electric components are compiled.

2.2.3 Examples for the Application of Ohm's Law

2.2.3.1 Charging of a Capacitor

A capacitor of capacitance C will be charged up to a voltage U_0 by a battery in series with a resistance R (Fig. 2.10). At time $t = 0$ when the switch S_1 is closed, the voltage at the capacitor is zero $U(0) = 0$. Because $Q(t) = C \cdot U(t)$ the charging current $I(t)$ is

$$I(t) = \frac{U_0 - U(t)}{R} = \frac{U_0}{R} - \frac{Q(t)}{R \cdot C}. \tag{2.9}$$

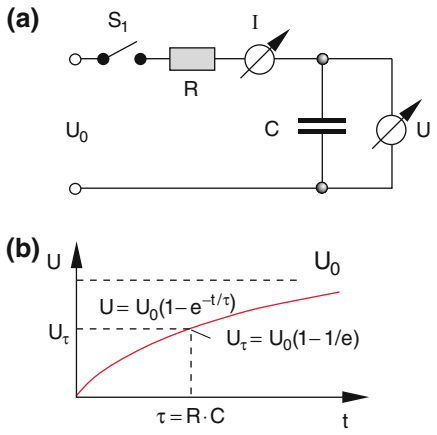


Fig. 2.10 Charging of a capacitor **a**) circuit **b**) voltage increase with time

Differentiating (2.9) with respect to time we get with $I(t) = \mathrm{d}Q/\mathrm{d}t$

$$\frac{\mathrm{d}I}{\mathrm{d}t} = -\frac{1}{R \cdot C} \cdot I(t),$$

By integration we obtain with the initial condition $I(0) = I_0$

$$I(t) = I_0 \cdot \mathrm{e}^{-t/(R \cdot C)}. \tag{2.10}$$

and with (2.9) the voltage across the capacitor

$$U(t) = U_0 \cdot \left(1 - \mathrm{e}^{-t/(R \cdot C)}\right). \tag{2.11}$$

2.2.3.2 Discharging a Capacitor

Now we consider the situation in Fig. 2.11 where the capacitor holds a voltage U_0 when the switch S_1 is closed and S_2 is open. At time $t = 0$ we close switch S_2 open S_1. Now a current $I(t)$ flows through the discharge resistance R_2

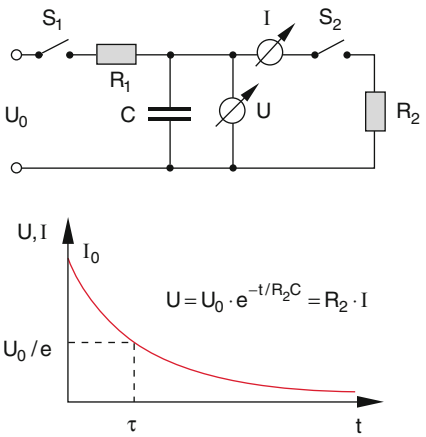


Fig. 2.11 Discharging of a capacitor

$$I(t) = -\frac{\mathrm{d}Q}{\mathrm{d}t} = -C \cdot \frac{\mathrm{d}U}{\mathrm{d}t} = \frac{U(t)}{R_2}. \tag{2.12}$$

The minus sign indicates that the charge on the capacitor plates decreases. Integration of (2.12) yields

$$U(t) = U_0 \cdot \mathrm{e}^{-t/(R_2 C)} \tag{2.13a}$$

and

$$I(t) = I_0 \cdot \mathrm{e}^{-t/(R_2 C)}. \tag{2.13b}$$

2.2.3.3 Voltage Divider

The constant voltage gradient across a conductor with current I can be used to supply a variable voltage $U < U_0$ from the constant source voltage U_0. As we see from Fig. 2.8 a variable voltage between point A and the sliding contact

$$U_1(x) = \frac{x}{L} \cdot U_0 \tag{2.14}$$

respectively

$$U_2(x) = \frac{x - L}{L} U_0$$

between point B and the sliding contact can be taped.

The practical realization of such a potentiometer consists of a thin conducting coating on a cylindrical insulator with a contact sliding over the coating to tap the variable voltage.

2.2.3.4 Resistance of a Flat Circular Ring of Thickness *h*

When we apply a voltage U across the inner cylinder (radius r_1) and the outer cylinder (radius r_2) (Fig. 2.12). then a current I flows radially through the dashed surface $A = 2\pi \cdot r \cdot h$ of a cylinder with thickness h and radius r between r_1 and r_2. We get

$$\begin{aligned} I &= \int \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A} = \sigma_{\mathrm{el}} \cdot \int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{A} \\ &= \sigma_{\mathrm{el}} \cdot E \cdot 2\pi \cdot r \cdot h. \end{aligned}$$

Because $\boldsymbol{E} = -\mathrm{grad}\phi = -\frac{\mathrm{d}\phi}{\mathrm{d}r}\hat{\boldsymbol{e}}_r$ it follows:

$$\begin{aligned} -\frac{\mathrm{d}\phi}{\mathrm{d}r} &= \frac{I}{2\pi \cdot \sigma_{\mathrm{el}} \cdot r \cdot h}, \\ \Rightarrow U = \phi_1 - \phi_2 &= \frac{I}{2\pi \cdot \sigma_{\mathrm{el}} \cdot h} \int_{r_1}^{r_2} \frac{\mathrm{d}r}{r} \\ &= \frac{I}{2\pi \cdot \sigma_{\mathrm{el}} \cdot h} \ln \frac{r_2}{r_1} \\ \Rightarrow R = U/I &= \frac{\ln(r_2/r_1)}{2\pi h \sigma_{\mathrm{el}}}. \end{aligned} \tag{2.15}$$

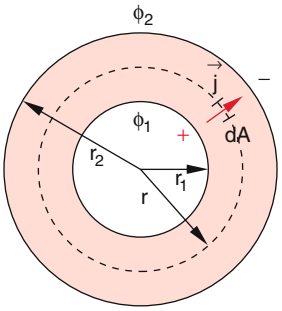


Fig. 2.12 Resistance of a flat circular ring between two concentric electrodes with radii r_1 and r_2

2.2.4 Temperature Dependence of the Electrical Resistance of Solids: Super-conductivity

If an electron collides with lattice atoms of a regular crystal, the momentum and collision energy is not transferred to a single atom but to the whole lattice. The reason is that each atom is bound elastically to its neighbors. Therefore, the electrons stimulate the whole crystal to vibrations that represent standing waves in the crystal and are called *phonons*. Boundary conditions select a finite number of possible vibrations that have discrete amounts of vibration energies and momenta. Because of these boundary conditions the wavelength of the standing waves cannot be smaller than twice the distance between the lattice planes and not greater than twice the size of the crystal. Therefore a discrete number of oscillation modes with discrete energies and momenta is excited. Electrons colliding with the lattice can create phonons and pass energy and momentum to the crystal.

In addition to its regular lattice atoms each real solid material has defects that are lattice points introduced by doping, where either atoms are missing, or extra atoms are sitting between the regular lattice points (see Vol. 3). These defects are not bound to each other in the same way as regular lattice atoms, and therefore can accept energy and momentum while they collide with electrons without stimulating lattice vibrations. The free path length Λ and thus the conductivity σ_{el} of metals become larger for pure materials. The specific resistance $\varrho_{\mathrm{s}} = 1/\sigma_{\mathrm{el}}$ can be composed of two terms

$$\varrho_{\mathrm{s}} = \varrho_{\mathrm{Ph}} + \varrho_{\mathrm{i}},$$

where ϱ_{Ph} is determined by interactions between electrons and phonons and ϱ_{i} by collision with impurity atoms or dislocations.

Examples

Electrons in copper at room temperature have a mean collision time $\tau_s = m \cdot \sigma_{el}/(n \cdot e^2) = 2.5 \times 10^{-14}$ s. For a mean velocity of 1.5×10^6 m/s the mean free path is $\Lambda = 4 \times 10^{-8}$ m = 40 nm. That corresponds to about 200 atomic diameters. The conductivity becomes, according to (2.6b), $\sigma_{el} = 6 \times 10^7$ A/Vm and the specific resistance is $\varrho_s = 1.7 \times 10^{-8}$ Vm/A. Commercially available copper is a polycrystalline material. Therefore, the contribution ϱ_i to the specific resistance dominates because the micro crystals are randomly oriented and there are no uniform phonons for the whole body.

2.2.4.1 Temperature Dependence of the Specific Resistance of Metals

The mean thermal velocity of the electrons rises with increasing temperature; in addition their free path Λ becomes smaller because more lattice vibrations are thermally excited.

Therefore the probability increases that the electrons can transfer energy and momentum to the phonons. Both effects cause a decrease of the electric conductivity $\sigma_{el}(T)$ and an increase of the specific resistance $\varrho_s(T) = 1/\sigma_{el}(T)$ of metals. This dependence can be written as

$$\varrho_s(T) = \varrho_0 \cdot (1 + \alpha \cdot T + \beta \cdot T^2) \qquad (2.16)$$

which is valid in a wide temperature range. In this equation is $\beta \cdot T \ll \alpha$. Within the for practical applications usually relevant limited temperature range $T_1 < T < T_2$ we can use the approximation

$$\varrho_s(T) \approx \varrho_0(1 + \alpha(T_m)T)$$

with a temperature dependent value of a and the mean temperature $T_m = (T_1 + T_2)/2$ of the chosen range. Table 2.2 lists ϱ_0 and α of a few metals. Figure 2.13 shows examples of their temperature dependence.

Table 2.2 Temperature dependence of the electric resistance $\rho(T_C) = \rho_0\,(1 + \alpha T_C)$ for some metals with $\rho_0 = \rho(T = 0$ °Celsius)

Metall	$\varrho_0/10^{-6}\,\Omega$m	α/K^{-1}
Silver	0.015	4×10^{-3}
Copper	0.016	4×10^{-3}
Aluminum	0.026	4.7×10^{-3}
Mercury	0.941	1×10^{-3}
Constantan ($Ni_{0,4}Cu_{0,5}Zn_{0,1}$)	0.5	$<10^{-4}$
Tungsten	0.05	4.83×10^{-3}

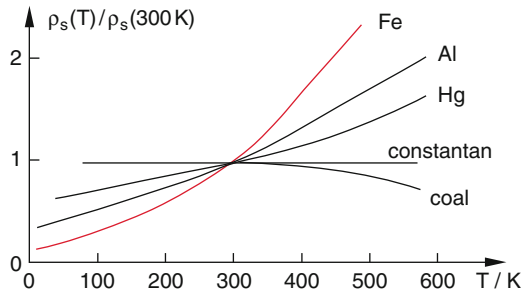


Fig. 2.13 Temperature dependence of the specific resistance of some conductors

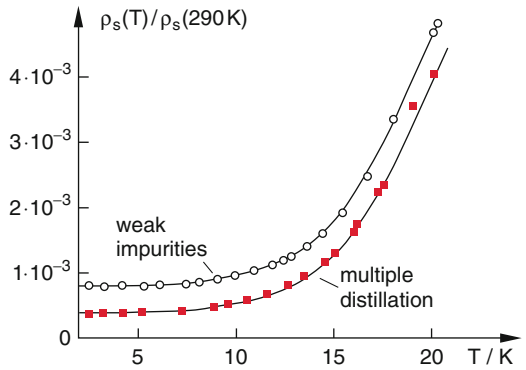


Fig. 2.14 Temperature dependence of two sodium samples with different concentrations of impurities

The number of excited lattice vibrations decreases with decreasing temperature. At very low temperatures it becomes very small and therefore the specific resistance should go in the limit $T - 0$ to a constant value which is caused by the impurities. Figure 2.14 shows the temperature dependence of two Na-samples with different percentages of impurities. A similar behavior can be found for a great number of metals. However, a few solid materials show a sudden vanishing of their dc resistance at a temperature T_C (superconductivity, Fig. 2.15).

2.2.4.2 Superconductivity

Superconductivity has been discovered for the first time 1911 by H. *Kamerlingh-Onnes* in Leiden, while he investigated the temperature dependence of the specific resistance of mercury (Hg) and its dependence on impurities. He cooled mercury which could be purified easily by repeated distillation, down to a temperature of 4 K.

These low temperatures were achieved by liquefying of helium. To his great surprise Kamerlingh-Onnes found that the resistance of his samples dropped suddenly down to zero at temperatures below 4.2 K. He named this phenomenon **superconductivity**. Later on the vanishing of the electric resistance had been found for other materials with different transition temperatures T_C [1]. Because of the technical

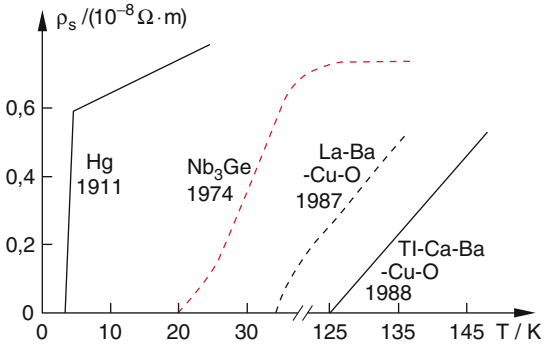


Fig. 2.15 Superconductivity of some conductors with different transition temperatures

importance of this discovery many investigations have been performed to find superconductors with a higher transition temperature. However, for several decades the transition temperatures remained below 30 K where liquid helium had to be used (Table 2.3).

Only in 1986 *Müller and Bednorz* at IBM (Rüschlikon, Swiss) found special oxide ceramics that became superconducting at transition temperatures above 80 K and therefore could be cooled by liquid nitrogen [2]. In 1987 both scientists got the Nobel Price for their discovery just as Kamerlingh-Onnes in 1913. In the meantime further superconducting oxides have been found (*high temperature superconductors*) that have transition temperatures above 120 K and therefore we are optimistic about technical applications in the nearer future [3].

It took also several decades to find a theoretical description of the superconducting state. Not until about 40 years after its discovery *Bardeen, Cooper and Schrieffer* developed a model that could explain most of the experimental facts. It is named BCS-theory after the initials of the three scientists. In this model the conduction electrons are kept together in pairs by a polarizing interaction with the lattice and become **Cooper-pairs**. These Cooper-pairs have a finite binding energy ΔE and therefore can only be separated if this energy is supplied by interactions with the lattice i.e. by collisions with the vibrating lattice atoms [1, 4–6].

This BCS theory can be illustrated by a simple mechanical model (Fig. 2.16): Two balls lie at different locations on the top of a rubber-membrane that is bulged in by the weight of

Table 2.3 Transition temperatures *Tc* for some superconductors

Element	T_c/K	Compound	T_c/K
Al	1.17	Al_2CMo_3	10.0
Hg	4.15	InNbSn	18.1
La	6.0	$AlGeNb_3$	20.7
Nb	9.25	LaBaCuO	85
		Tl-Ca-Ba-CuO	125

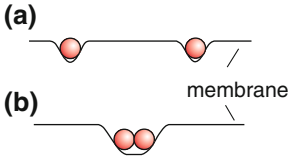


Fig. 2.16 Simple model of a Cooper-Pair

the balls. If the balls are brought together then the dent in the membrane becomes deeper because the weight at this position is now twice as large. The potential energy of the balls becomes lower than in the case of separation i.e. the elastic membrane provides a binding energy between the balls.

We have to supply energy to separate the balls again.

Applied to the Cooper pairs this model tells us: Each electron polarizes the electron shell of the ions by its Coulomb interaction with the lattice ions during its motion though the lattice (Fig. 2.17). If a second electron with opposite but equal momentum moves along the same track though the lattice it suffers besides its Coulomb interaction with the ion cores an additional attractive interaction between its charge and the induced charge polarization of the lattice, created by the first electron. Also the first electron experiences the additional attractive interaction caused by the second electron. As a result the potential energy of both electrons is reduced and that of the lattice is raised.

We say: The polarization of the lattice introduces a correlation between both electrons that yields a "bound" electron pair $(e^-, \boldsymbol{p};\ e^-, -\boldsymbol{p})$ with the momenta $+\boldsymbol{p}$ and $-\boldsymbol{p}$ of the two electrons and the total momentum $\boldsymbol{p}_C = +\boldsymbol{p} + (-\boldsymbol{p}) = \boldsymbol{0}$. Because without an external field the total momentum of the Cooper pair is zero it cannot supply any kinetic energy to the lattice as long as the thermal energy of the phonons interacting with the Cooper pair is lower than its binding energy. If, however, an electric field is applied the drift velocity v_D superimposes the velocity v of both pair electrons and the momentum of the Cooper pair becomes

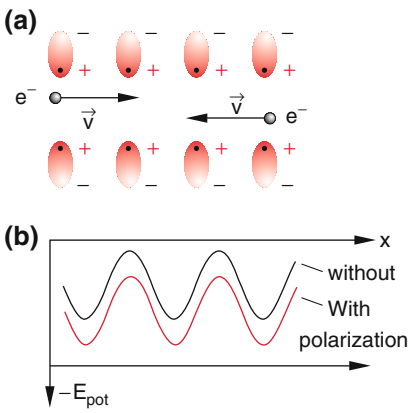


Fig. 2.17 Polarization of lattice ions by electrons flying through the lattice. **a**) Schematic illustration **b**) effective potential with and without polarization, experienced by an electron

$$\boldsymbol{p} = 2m_e \cdot \boldsymbol{v}_\mathrm{D}.$$

This results in a current density

$$\boldsymbol{j} = 2 \cdot n_\mathrm{C} \cdot \mathrm{e} \cdot \boldsymbol{v}_\mathrm{D},$$

where n_C is the number density of Cooper pairs. Now the friction force is missing because the Cooper pair cannot transfer energy to the lattice and the current remains constant even if the field is switched off. It has been shown that the superconducting current did not measurably change over a whole year.

If the external electric field is not switched off the drift velocity v_D resp. the electric current rises until the additional kinetic energy of the Cooper pair

$$\begin{aligned}\Delta E_\mathrm{kin} &= \frac{1}{2}(2m)(\boldsymbol{v}+\boldsymbol{v}_\mathrm{D})^2 - 2\cdot\frac{1}{2}mv^2 \\ &= 2m\boldsymbol{v}\cdot\boldsymbol{v}_\mathrm{D} + mv_\mathrm{D}^2\end{aligned}$$

becomes larger than the negative binding energy. Then the Cooper pair decays into two normal electrons which again interact with the lattice and therefore their drift velocity becomes lower than that of the Cooper pair. The superconductivity passes over into normal conductivity.

Also without an external field the Cooper pairs decay above the critical temperature T_C when the additional thermal energy becomes larger than the binding energy.

Although this model of the BCS theory for Cooper pairs can explain correctly many experimental facts there are still a series of observations for which up to now no satisfactory explanation can be given. Especially the new high temperature superconductors cannot be described by the Cooper pair model. Here new theoretical starting points have proved that take into account the layered structure of the materials (*perovskite*) and the resulting directional dependent conductivity (see Vol. 3 and [5].).

2.2.4.3 Conductivity of Semiconductors and Its Variation with Temperature

The situation is different for semiconductors. Their conductivity is determined mainly by the number density n of free conduction electrons.

In pure semiconductors at room temperature the number density of free electrons and thus the conductivity are very low but they can increase by many orders of magnitude by adding impurity atoms (see Vol. 3). This can be seen in Fig. 2.18 where the specific resistances $\varrho_\mathrm{s}(T, n_\mathrm{D})$ for various degrees of doping n_D are plotted against the reciprocal temperature.

Note the logarithmic scale for resistivity and conductivity.

The density n of the free electrons increases exponentially with temperature

$$n(T) = n_0 \cdot \mathrm{e}^{-\Delta E/kT}$$

where ΔE is the energy that has to be supplied in order to transfer electrons from the bound state to free conduction electrons.

In doped semiconductors impurity atoms have been brought into the crystal which have much smaller energies ΔE. Therefore at low temperatures mainly the impurity atoms (donors) provide the conduction electrons. Above a saturation temperature T_s all donors are ionized and the number of charge carriers does not further increase because the contribution of the crystal atoms which also rises with T is still very small. As the mobility u decreases with increasing temperature the conductivity σ_el declines above T_s again (Fig. 2.19).

In the temperature range below T_s the decrease of the mean free path $\Lambda(T)$ and thus the mobility $u(T)$ is overcompensated with increasing T by the steep rise of the density $n(T)$ of the conduction electrons. Therefore the conductivity $\sigma_\mathrm{el}(T)$ increases in this range with increasing temperature i.e. the specific resistance $\varrho_\mathrm{s}(T)$ decreases.

Altogether, in this temperature range semiconductors have a negative temperature coefficient $\alpha = [\mathrm{d}\varrho/\mathrm{d}T]/\varrho_0$ of their

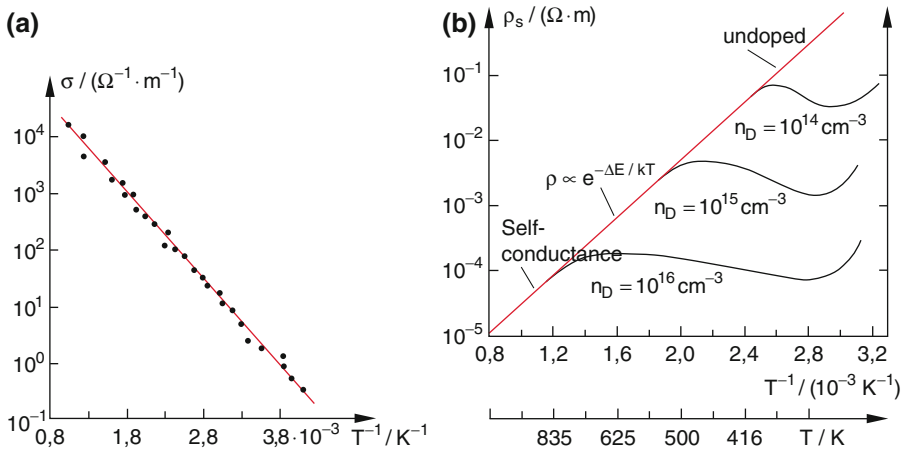


Fig. 2.18 **a**) Electrical conductivity of a semiconductor as a function of temperature **b**) specific resistance of semiconductors with different doping concentrations

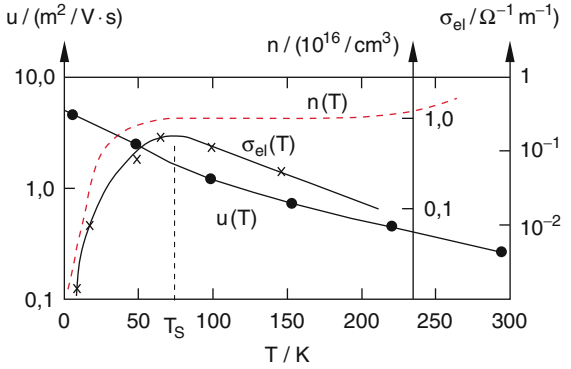


Fig. 2.19 Electrical conductivity, electron density *n(T)* and mobility *u(T)* of doped semiconductors as a function of temperature

Table 2.4 Comparison of the temperature dependence of the relative specific resistance $\rho_s(T)/\rho_s$ (0 °C). For a metal (Cu) and a semiconductor (Ge)

$\varrho_s(T)/\varrho_s(T = 273\,\mathrm{K})$	Cu	Ge
273	1	1
300	1.12	0.8
400	1.55	1.2×10^{-2}
500	1.99	1.4×10^{-3}
600	2.43	3×10^{-4}
800	3.26	8×10^{-5}
1000	4.64	–

specific resistance and are therefore called NTC-resistances (negative temperature coefficient).

In Table 2.4 specific values of a metal (copper) and a semiconductor (germanium) are listed.

The strong dependence of the resistance of semiconductors on temperature is used to build sensitive temperature sensors. If such a semiconductor is used in a voltage divider (Fig. 2.20) every temperature variation yields a variation of the voltage U at the output terminals and a signal indicating the temperature. It can be also used to set up a circuit for controlling temperatures.

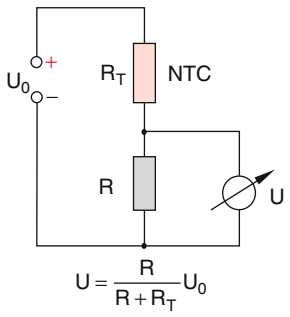


Fig. 2.20 Electric circuit for temperature measurements with an NTC-semiconductor resistance

2.3 Electric Power and Joule's Heating

The transport of a charge q from a position with potential ϕ_1 to a position with potential ϕ_2 requires the work (see (1.13))

$$W = q \cdot (\phi_1 - \phi_2) = q \cdot U$$

For a constant voltage $U = \phi_1 - \phi_2$ the charge $\mathrm{d}Q/\mathrm{d}t$ flowing per unit time through a conductor yields the electric power

$$P = \frac{\mathrm{d}W}{\mathrm{d}t} = U \cdot \frac{\mathrm{d}Q}{\mathrm{d}t} = U \cdot I, \qquad (2.17a)$$

Its unit is $[P] = \mathrm{V} \cdot \mathrm{A} = \mathrm{Watt} = \mathrm{W}$.

The work supplied during the time interval $\Delta t = t_2 - t_1$ is

$$W = \int_{t_1}^{t_2} U \cdot I \mathrm{d}t = U \cdot I \cdot \Delta t, \qquad (2.17b)$$

The second equal sign is only valid if U and I are temporally constant. The unit of work is watt second = Ws = Joule = J = Nm

This electric energy is transformed into heat by the frictional force $\boldsymbol{F}_R = -k_R \cdot \boldsymbol{v}_D$ that is opposite and equal to the force $q \cdot \boldsymbol{E}$. The conductor becomes hot (Joule's heat).

For Ohmic conductors that obey Ohm's law $U = I \cdot R$ we can express the electric power as

$$P = U \cdot I = I^2 \cdot R = \frac{U^2}{R}, \qquad (2.18)$$

Figure 2.21 demonstrates that those parts of a conductor with the higher resistance consume the higher power if the

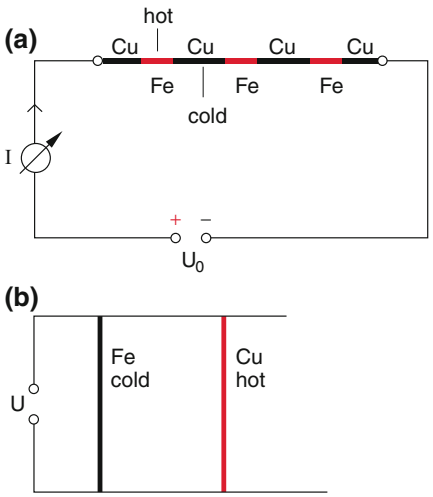


Fig. 2.21 Demonstration of electric power transformed into heat **a**) for constant current **b**) for constant voltage

current through the conductor is constant. However, if the voltage across each resistance has the same value the consumed power of the resistor rises with decreasing resistance!

2.4 Electric Networks; Kirchhoff's Rules

Electric circuits often consist of a network of conductors with nodes or junctions where currents meet or drain away. We will use the following rules to calculate the individual currents, voltages and the total resistance.

Kirchhoffs first rule (junction rule)

If conductors meet at one point P the sum of all currents entering the junction must equal the sum of all currents leaving the junction i.e. the algebraic sum of all currents must be zero (Fig. 2.22).

$$\sum_k \boldsymbol{I}_k = 0 \tag{2.19}$$

This result follows from the continuity equation because at the point P charges are neither created nor destroyed and so the total current through a closed surface $\boldsymbol{A}$ around point P must be zero. According to (2.4a) is

$$\begin{aligned} -\frac{\mathrm{d}Q}{\mathrm{d}t} &= -\frac{\mathrm{d}}{\mathrm{d}t}\int_V \varrho_{\mathrm{el}} \cdot \mathrm{d}V \\ &= \int_V \mathrm{div}\,\boldsymbol{j}\,\mathrm{d}V \\ &= \int_A \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A} = \sum_k I_k = 0. \end{aligned}$$

Kirchhoffs second rule (loop rule)

The algebraic sum of the voltages across elements in a closed circuit path (loop) must be equal to the generator voltage (Fig. 2.23).

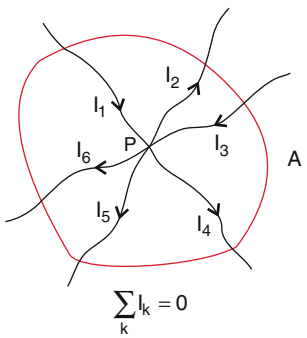


Fig. 2.22 Illustration of Kirchoff's junction rule

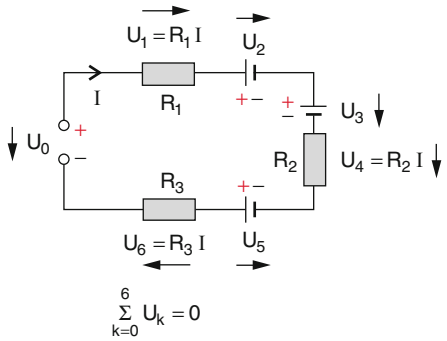


Fig. 2.23 Illustration of Kirchoff's loop rule

$$U_0 = \sum_{k=1}^{N} U_k = R_1 I + U_2 + U_3 + R_2 I - U_5 + R_3 I, \tag{2.20a}$$

where the sum extends over all voltage sources and all consumers. In Fig. 2.23 for example this gives

$$\mathrm{U}_1 = \mathrm{R}_1 \cdot \mathrm{I};\ \mathrm{U}_4 = \mathrm{R}_2 \cdot \mathrm{I};\ \mathrm{U}_6 = \mathrm{R}_3 \cdot \mathrm{I},$$

whereas the internal voltage sources are $U_2 + U_3 - U_5$. The external voltage source is U_0.

Kirchhoff's rule then requires:

$$U_0 = R_1 \cdot I + U_2 + U_3 + R_2 \cdot I - U_5 + R_3 \cdot I$$

Kirchhoff's rule is also valid for capacitive or inductive resistances (see Sect. 5.4)

$$\sum_{k=0}^{N} U_k = 0. \tag{2.20b}$$

2.4.1 Resistances in Series

If we connect several resistances R_k in series to the circuit in Fig. 2.24 with a current I then the voltage drops at the resistances are

$$U_k = I \cdot R_k$$

From (2.20a) it follows

$$U_0 = \sum U_k$$

The total resistance $R = \sum R_k$ is equal to the sum of the individual resistances.

Resistors arranged in series (one behind the other) add to a total resistance.

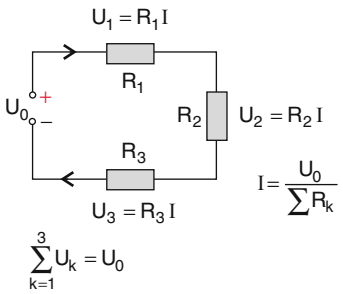


Fig. 2.24 Circuit with 3 resistors in series

2.4.2 Parallel Arrangement of Resistors

If two resistances are connected in parallel (Fig. 2.25) the total resistance is smaller than the individual resistances.

When we apply a voltage U between the points A and B (Fig. 2.25), we get the condition

$$\frac{U}{R} = I = I_1 + I_2 = \frac{U}{R_1} + \frac{U}{R_2} \Rightarrow \frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2}. \tag{2.21}$$

For the parallel connection of resistors the inverse of the individual resistances add up to the inverse of the total resistance.

Therefore, the total resistance R is smaller than the smallest value of the two resistances

By using the conductivity $G = 1/R$ (2.21) simplifies to

$$G = G_1 + G_2. \tag{2.21a}$$

We can formulate the rules for parallel and series networks of resistors:

For resistances placed in series (one behind the other) the individual resistances add to the total resistance.

For resistances placed parallel to each other the individual conductivities add to the total conductivity.

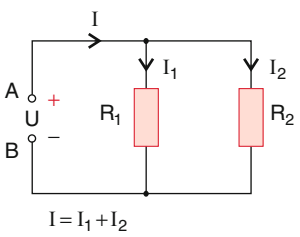


Fig. 2.25 Circuit with two parallel resistors

2.4.3 The Wheatstone Bridge

To measure resistances very accurately a Wheatstone bridge can be used (Fig. 2.26). The values of R_1, R_2, and R_3 are known, but R_x is unknown. Across the points A and B a voltage U_0 is applied. The voltages

$$U_1 = U_0 \cdot \frac{R_x}{(R_1 + R_x)} \quad \text{and} \quad U_2 = U_0 \cdot \frac{R_2}{(R_2 + R_3)}$$

at the points C and D referenced to B are equal if

$$\frac{R_1}{R_x} = \frac{R_3}{R_2} \Rightarrow U_1 = U_2.$$

Then the voltage $\Delta U = U_1 - U_2$ between the points C and D is zero, i.e. the current through the measuring instrument is zero and it follows

$$R_x = \frac{R_1 \cdot R_2}{R_3}.$$

Usually a variable voltage divider (potentiometer) is used to adjust the bridge. So R_2 and R_3 can be varied simultaneously (Fig. 2.26). If the voltage divider has a length of L and the wiper contact is at the position x we get

$$\frac{R_2}{R_3} = \frac{L - x}{x}.$$

The unknown resistance is then

$$R_x = R_1 \frac{L - x}{x}. \tag{2.22}$$

The adjustment to zero is very sensitive because the measuring instrument can detect very low currents and voltages $U_1 - U_2$. Therefore, the Wheatstone bridge provides a means to measure precisely resistances and their temperature dependence. The essential point is that the adjustment to zero is independent of the applied voltage U_0.

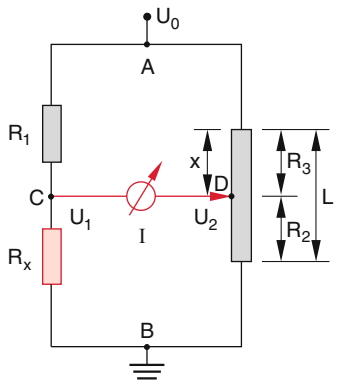


Fig. 2.26 Wheatstone Bridge

2.5 Methods to Measure Electric Currents and Voltages

In principle, all effects that are caused by electric currents can be used to measure electric currents. These are especially Joule's heat, magnetic effects, electrolytic dissociation of conducting fluids and the voltage drop in conductors. A device for the measurement of electric currents is named **ampere-meter**. Some of these instruments will be explained in more detail (see also [6–8]).

2.5.1 Current Measuring Instruments

2.5.1.1 Hot Wire Ampere-Meter

If a current I passes through a wire with resistance R the electric power $P = RI^2$ is transformed into heat and the wire heats up. That causes an expansion of the wire (see Vol. 1, Chap. 10). In a hot wire ampere-meter a system of levers transfers the change of length to a revolving pointer. A spring keeps the lengthened hot wire always tightened (Fig. 2.27). These instruments are robust but insensitive; they are suited for currents higher than about 0.1 A.

2.5.1.2 Using Magnetic Effects to Measure Currents

Electric currents create magnetic fields (see Chap. 3) which exert forces or torques on magnetic dipoles. This is used to bring about the mechanical motion of a pointer.

In a moving coil instrument (Fig. 2.28) a coil, carrying the current to be measured, is connected to a pointer. The coil can revolve in a magnetic field. The coil experiences a torque which is proportional to the current through the coil and turns the pointer against the restoring force of the spiral spring (see Sect. 3.5.1). Instruments that use the interaction

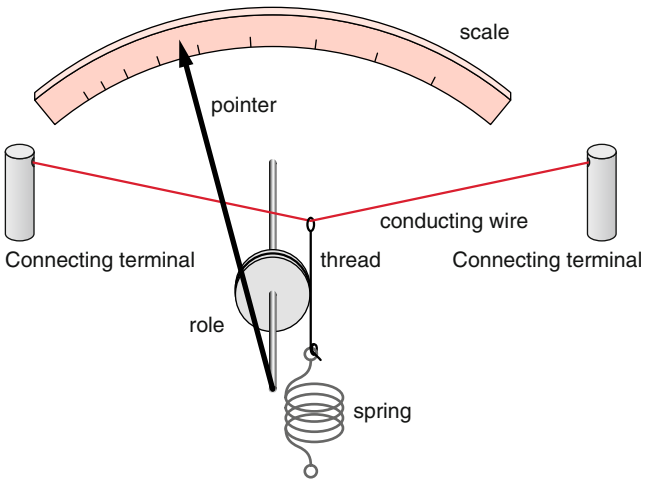


Fig. 2.27 Hot wire ampere-meter

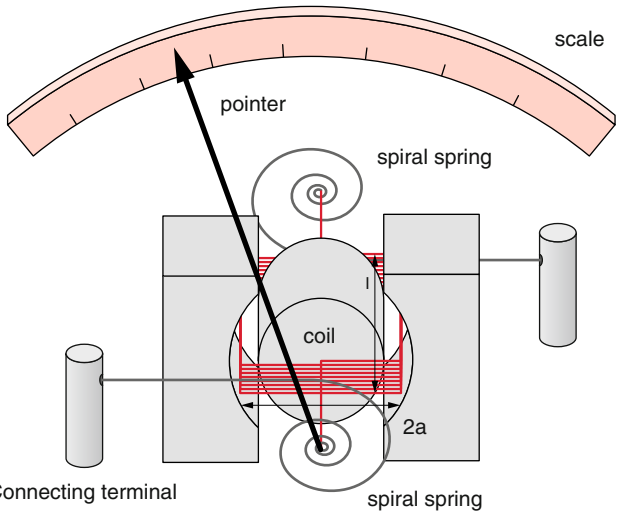


Fig. 2.28 Moving coil ampere-meter (galvanometer)

between the current flowing in a coil and magnetic fields are called *galvanometers* or *moving coil instruments*.

In a moving iron instrument (Fig. 2.29a) the current carrying coil is fixed and a soft iron bar inside the coil becomes magnetized and experiences a torque which is proportional to the current through the coil. The pointer is connected to the axis of the bar and indicates the current I. In the soft iron

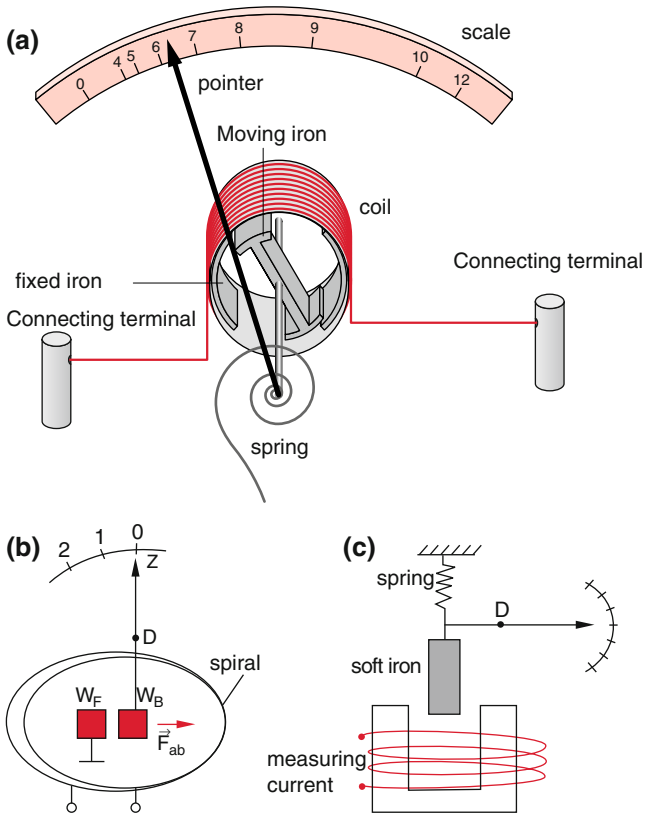


Fig. 2.29 Moving iron ampere-meter **a**) schematic drawing **b**) illustration of repelling iron bars **c**) soft iron instrument

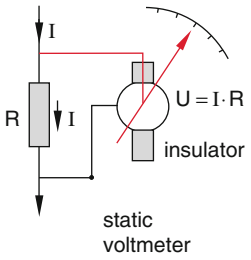


Fig. 2.30 Measuring current with a static voltmeter

instrument (Fig. 2.29b) the current to be measured flows through a coil and creates a magnetic field that magnetizes two rods of soft iron in the same direction so that they repel each other. If the current changes its direction, also the direction of magnetization is changed for both pieces of iron and again they repel. The reading is therefore independent of the direction of the current flow. Such instruments can be also used to measure alternating currents. A modification of this method uses one rod of iron that is pulled into the magnetic field created by the current to be measured (Fig. 2.29c).

Here *soft iron* is used because this material can be readily magnetized in one direction but its magnetization can be also easily turned into the opposite direction when the magnetic field is commutated. The hysteresis loop encloses only a small area (see Sect. 3.5.5).

2.5.1.3 Electrolytic Effects to Measure Currents

Many liquid molecular materials are chemically decomposed if they transport an electric current (Sect. 2.6). The molecules dissociate into positive and negative ions, which are deposited at the electrodes. The mass deposited per unit time is proportional to the current and therefore can be used for measuring the current.

2.5.1.4 Measurement of Currents with a Static Voltmeter

When the current I flows through a resistor R it generates a voltage drop $U = I \cdot R$ and therefore the current I can be, in principle, measured by a static voltmeter connected parallel to R if its internal resistance is large compared to R (Fig. 2.30).

2.5.2 Circuits with Ampere-Meters

Each ampere-meter has a full scale value for the current to be measured depending on the construction of the instrument. This range can be extended to higher currents by a resistance in parallel to the ampere-meter (Fig. 2.31). For an internal resistance R_i of the instrument and a resistance R parallel to it only the fraction $I_i = I \cdot R/(R + R_i)$ of the total current flows through the instrument.

The measurement of the current I by an instrument with total resistance $R_M = R \cdot R_i/(R + R_i)$ changes the original

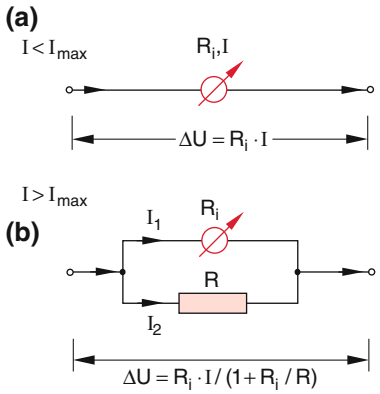


Fig. 2.31 Measuring currents with one of the instruments shown before **a**) if the current is smaller than the maximum allowed current through the instrument **b**) if $I > I_{max}$

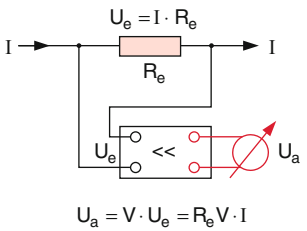


Fig. 2.32 Sensitive measurements of small currents by amplification of the voltage $U_e \cdot R_e$

voltage in the circuit because a voltage drop of $\Delta U = R_M \cdot I$ appears across the measuring instrument. Therefore the resistance R_M of an ampere-meter should be as small as possible. This can be achieved with instruments of high sensitivity which can still measure very low currents I_i.

Modern instruments amplify the voltage drop $U_e = R_e \cdot I$ at the input resistance R_e by a factor V (Fig. 2.32). This allows one to measure currents down to 10^{-16} A.

Example

$I = 10^{-10}\text{A}, R_e = 10\,\text{k}\Omega \Rightarrow U_e = 1\,\Omega\text{V}, U_a = V \cdot U_e = 1\,\text{V}$ if $V = 10^6$.

2.5.3 Current Measuring Instruments Used to Measure Voltages

A voltage U across a resistance R causes a current $I = U/R$ through R and therefore we can use an ampere-meter to measure voltages.

A resistor R is used in series with the measuring instrument which has the inner resistance R_i (Fig. 2.32). The external resistor R is chosen such that the current

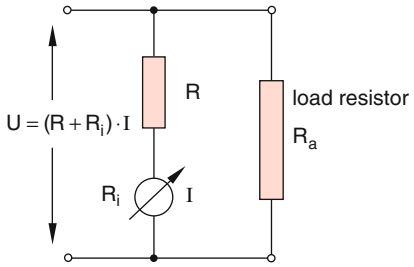


Fig. 2.33 Use of an amperemeter to measure voltages

$I = U/(R + R_i)$ lies in the measurement range of the instrument. Current measuring instruments that are used to measure voltages must have a very high total resistance $R + R_i$ in order not to influence the currents in the circuit (Fig. 2.33).

Ampere-meters are supposed to have a very low total resistance but voltmeters a very high one. Voltages and currents can be measured with the same instrument. For current measurements a high resistance must be placed parallel to the instrument, whereas for voltage measurements a high resistance has to be placed in series with the instrument.

2.6 Ionic Conduction in Fluids

Between two metallic electrodes inserted into a fluid where acids, alkaline solutions or salts are added (Fig. 2.34) a current I flows if a voltage U is applied which generates an electric field E between the electrodes. Such electrically conducting fluids are called **electrolytes**. In contrast to the metallic conductors the flow of an electric current through the fluid is connected with a chemical decomposition of the electrolyte. At the positive electrode (anode) as well as at the negative electrode (cathode) (Fig. 2.34) material in the solid or gaseous phase is deposited.

For example for a copper-sulfate solution in water $CuSO_4$ molecules dissociate into positively charged Cu^{++}-ions and negatively charged SO_4^{--}-ions even without an applied voltage because of their interactions with the water molecules. Etymologically, the word ion is derived from the Greek meaning "hiking". A voltage across the electrodes creates an electric field that moves the positive ions (*cations*)

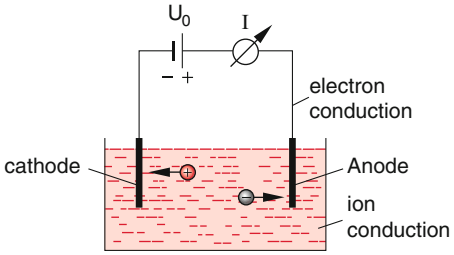


Fig. 2.34 Electric conduction in a galvanic cell

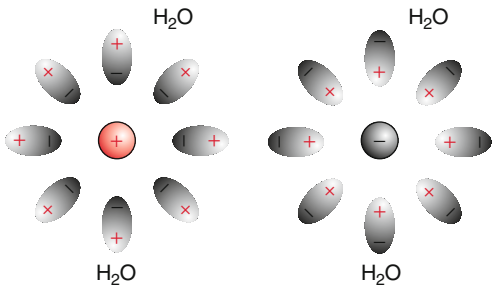


Fig. 2.35 Attraction of electric dipole water molecules by a positive or negative ion

to the cathode where the Cu^{++}-ion takes two electrons from the cathode, where it is deposited as neutral metallic copper. The negative ions (*anions*) move to the anode. There they release two electrons and the neutral SO_4 molecule reacts with water as

$$2SO_4 + 2H_2O \rightarrow 2H_2SO_4 + O_2.$$

and oxygen escapes at the anode as a gas.

All electrolytes are composed of molecules with non-symmetric distribution of electrons (the center of the positive charge distribution does not coincide with the center of the negative charge distribution and the molecules have a permanent dipole moment). These permanent dipoles are attracted by ions as is illustrated in Fig. 2.35.

In an electrolytic solution such molecules dissociate into ions of opposite signs. The energy ΔW_1 is necessary to dissociate them. When the ions attach themselves to the water molecules energy is released.

$$\begin{aligned} CuSO_4 &\xrightarrow[+\Delta W_1]{} Cu^{++} + SO_4^{--} \\ Cu^{++} + nH_2O &\xrightarrow[-\Delta W_2]{} \left(Cu \cdot nH_2O\right)^{++} \\ SO_4^{--} + nH_2O &\xrightarrow[-\Delta W_3]{} \left(SO_4 \cdot nH_2O\right)^{++} . \end{aligned}$$

The dissociation of the electrolyte molecules in water into pairs of ions occurs spontaneously if the gain in energy $\Delta W_2 + \Delta W_3$ for the attachment of the ions to the water molecules is larger than the dissociation energy ΔW_1.

When starting with small values of the concentration (n molecules/m^3) of salt molecules dissolved in water the electric current I at a constant voltage U increases at first with increasing concentration (Fig. 2.36). The conductivity increases linearly with the concentration until saturation starts, reaches a maximum and declines again at high concentrations.

This can be understood as follows: According to Eq. (2.6c) the electric conductivity

$$\sigma_{el} = n \cdot q \cdot u$$

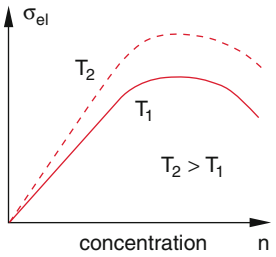


Fig. 2.36 Electrical conductivity of an electrolyte as a function of the concentration in water for two different temperatures

Table 2.5 Ion mobility in aqueous solutions at $T = 20$ °C for very small ion concentrations

Cations	u^+ m^2/(V s)	Anions	u^- m^2/(V s)
H^+	31.5×10^{-8}	OH^-	17.4×10^{-8}
Li^+	3.3×10^{-8}	Cl^-	6.9×10^{-8}
Na^+	4.3×10^{-8}	Br^-	6.7×10^{-8}
Ag^+	5.4×10^{-8}	I^-	6.7×10^{-8}
Zn^{++}	4.8×10^{-8}	SO_4^{--}	7.1×10^{-8}

can be written as the product of the ion concentration n and the mobility u. For low concentrations the mobility u is independent of n and its value is about $10^{-8} - 10^{-7}$ m^2/(Vs) (Table 2.5). In this concentration range the conductivity σ_{el} increases linearly with n.

Example

For $n^+ = n^- = 10^{24}$/m^3 (low concentration of ions, corresponding to a molar concentration of 1.5 mol/m^3), and the mobility $u^+ = 4.3 \times 10^{-8}$ m^2/Vs for Na^+ and $u^- = 6.9 \times 10^{-8}$ m^2/Vs for Cl^- ions the electric conductivity of a NaCl solution becomes $\sigma_{el} = (n^+ u^+ + n^- u^-) \cdot e = 1.8 \times 10^{-2}$ A/Vm. At an electric field $E = 10^3$ V/m the drift velocity becomes $v_D^+ = 4.3 \times 10^{-5}$ m/s and $v_D^- = 6.9 \times 10^{-5}$ m/s. This results in a current density $j = \sigma_{el} \cdot E = 18$ A/m^2.

With increasing concentration n the mean distance between the ions decreases and the attractions between the ions increases. Work has to be done to spatially separate the ions. This can be expressed by frictional forces which we have already discussed in Sect. 2.2. They increase with increasing concentration of ions due to the long range Coulomb forces F_C between the ions which are proportional to $1/r^2$ whereas they decrease with $1/r^6$ for interactions between neutral molecules.

Therefore the mobility decreases with increasing concentration at first slowly than faster and faster and the increase of n is overcompensated by the decrease of u.

The conductivity σ_{el} of electrolytes incrseases with increasing temperature in contrast to metals where it decreases. There are two reasons

(1) The viscosity of the solvents decreases with increasing temperature and therefore the mobility u increases.
(2) The thermal energy of the ions increases with T and less energy has to be supplied against the Coulomb attraction to separate the ions.

One mole of an ion with charge $Z \cdot e$ transports the charge

$$Q = N_A \cdot Z \cdot e = F \cdot Z$$

where N_A is the Avogadro constant (number of molecules per mole) and F is the Faraday constant .

The **Faraday constant** gives the charge that is transported by one mole of ions with one electron missing or one extra electron attached ($Z = 1$).

$$F = N_A \cdot e = 96\,485.309\,\text{C/mol}.$$

During the transport of the charge F a mass $m = M/Z$ is transported where M is the molar mass of the ions. The mass of the ions that is deposited at the electrodes while the charge of 1 C is transported is called electrochemical equivalent E_C.

Example

½ · 63.5 g Cu^{++}-ions transport the charge $F = 9.6 \times 10^4$ C i.e. for the charge transport of 1 C the mass of the cathode increases by $31.75/96\,000$ g $= 0.33$ mg.

Measuring the current I during the time Δt and the resulting change of mass Δm of the copper cathode yields the elementary charge $e = 1.6022 \times 10^{-19}$ C.

2.7 Current in Gases and Gas Discharges

Partially or totally ionized gases are called *plasma*. They belong to the mixed conductors, because the charges are transported by electrons as well as by positive and negative ions. Apart from a few exceptions a plasma is quasi-neutral because the mean number of negative and positive charges is equal in a volume of at least $\Delta V \approx r_D^3$. The quantity r_D is called **Debye-length.**

2.7.1 Concentration of Charge Carriers

The density of charge carriers $n^+ \approx n^- = n$ in a quasi-neutral plasma is determined by the creation rate $(dn/dt)_{erz} = \alpha$ and the rate of annihilation of the ion-pairs.

Recombination is the most important process of annihilation where an electron and a positive ion collide and form a neutral atom or molecule. The kinetic energy of their relative motion before the collision is either transformed into recombination radiation or transferred to a third collision partner which also can be the wall of the container.

The rate of recombination is proportional to the product $n^+ \cdot n^-$ of the densities of electrons and ions. It is therefore

$$(\mathrm{d}n/\mathrm{d}t)_{\mathrm{rek}} = -\beta \cdot n^2$$

Altogether the rate of change of the charge density is given by

$$\frac{\mathrm{d}n}{\mathrm{d}t} = \alpha - \beta n^2. \tag{2.23}$$

where α is the creation rate of ion pairs. Stationary equilibrium $\mathrm{d}n/\mathrm{d}t = 0$ is reached, if the rates of creation and of destruction are equal. Then the stationary density of charge carriers becomes

$$n_{\mathrm{stat}} = \sqrt{\alpha/\beta}. \tag{2.24}$$

Note The quantity n is the number density of ion-pairs. Therefore we have a total of $2n$ charge carriers ($n^+ + n^- = 2n$) per unit volume.

If the generation of electrons ends at time $t = 0$ with a density of charge carriers $n_0 = n(t = 0)$ then the number density $n(t)$ decreases by recombination. Integration of (2.23) with $\alpha = 0$ yields

$$n(t) = \frac{n_0}{1 + \beta n_0 t} = \frac{n_0}{1 + t/\tau_{1/2}}. \tag{2.25}$$

The decay curve $n(t)$ is a hyperbola. The half-life $\tau_{1/2} = 1/(\beta \cdot n_0)$ is the time during which the concentration decreases to half of its initial value n_0.

2.7.2 Creation of Charge Carriers

In gases pairs of ion-electron can be created in various ways.

2.7.2.1 Thermal Ionization

Positioning a burning candle or a Bunsen burner below the vertical electrodes of a parallel plate capacitor a current begins to flow in the circuit of battery and capacitor (Fig. 2.37). The current ends again if we remove the candle. Obviously the flame creates carriers of charges that are transported by the electric field and hit the-plates of the capacitor. The carriers of charges are created by a combination of thermal excitation or ionization and chemical reactions in the flame initiated by this ionization.

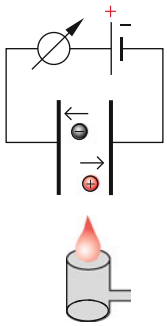


Fig. 2.37 Thermal ionization of the air in the electric field between the plates of a capacitor by a Bunsen burner

Example

At the temperature of the surface of the sun, $T = 6000$ K, only a fraction of 10^{-4} of the neutral hydrogen atoms is ionized.

Surfaces of special solid materials act as catalyst and can substantially increase the fraction of ionization at lower temperatures.

2.7.2.2 Ionization by Collisions with Electrons

An electron with enough kinetic energy, $E_{\mathrm{kin}} > E_{\mathrm{ion}} \approx 10$ eV colliding with an atom A or a molecule M with the ionization energy E_{ion} can knock out one electron of the atomic or molecular shell creating an electron-ion-pair

$$\begin{aligned} \mathrm{e}^- + \mathrm{A} &\rightarrow \mathrm{A}^+ + \mathrm{e}^- + \mathrm{e}^- \\ \mathrm{e}^- + \mathrm{M} &\rightarrow \mathrm{M}^+ + \mathrm{e}^- + \mathrm{e}^-. \end{aligned} \tag{2.26}$$

This is the main mechanism in gases of creating charge carriers.

2.7.2.3 Photo-ionization

If we irradiate the air between the plates of a charged capacitor with ultraviolet light of short wavelength or with X-rays then a current is generated proportional to the intensity of the radiation. The electron-ion pairs carrying this current are created by photo-ionization of the gas molecules according to

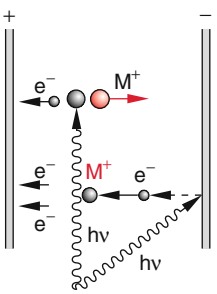


Fig. 2.38 Creation of ions by electron impact or by photo-ionization in the electric field of a parallel plate capacitor

$$M + h\nu \rightarrow M^+ + e^-.$$

If radiation hits the plates of the capacitor then electrons are released of the negative plate. In the electric field they gain sufficient energy to create new electron-ion-pairs by collision with air molecules (Fig. 2.38).

2.7.3 Current-Voltage-Characteristic of a Gas Discharge

When we create charge carriers by one of the methods described above inside a container filled with gas at a pressure of a few millibars and two electrodes A and K we can measure the functional dependence $I(U)$ of the discharge current I as a function of the voltage U between A and K (Fig. 2.39).

We start with a small voltage U. At the beginning the current $I(U)$ increases proportional to the voltage U. This is called the *linear range*. Increasing the voltage U further the current stays nearly constant at a value I_S that is independent of U. This is called the *saturation range*. Above a critical voltage U_C, which depends on the type of gas, the gas pressure and the geometry of the container, a steep increase of the current I occurs. Here the range of collision-ionization is reached.

Above this range at the ignition voltage U_Z a self-sustained discharge starts which continues also without external creation of charge carriers.

This current-voltage characteristic $I(U)$ can be explained as follows: The charge carriers created by one of the processes described above get a drift velocity due to the electric field E between K and A

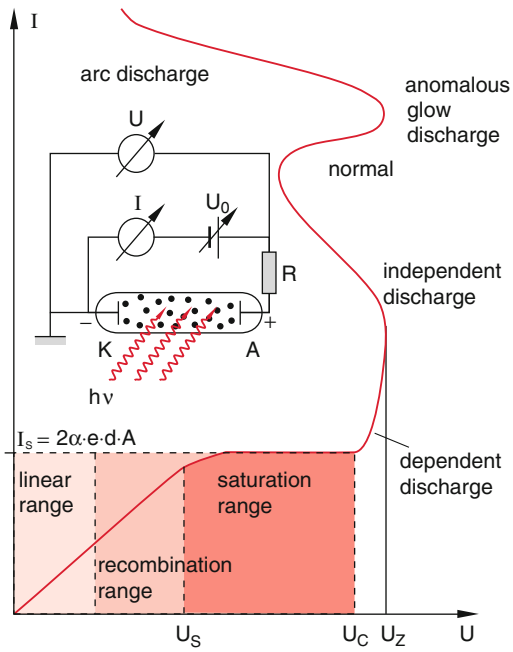


Fig. 2.39 Current-voltage characteristics of a gas discharge

$$\boldsymbol{v}_D = \frac{e \cdot \tau_s}{m}\boldsymbol{E},$$

This drift velocity depends on the field strength $\boldsymbol{E}$, on the mean collision time $\tau_s = \Lambda/\bar{v}$ and thus due to the mean free path $\Lambda = kT/(p \cdot \sigma_c)$ on the pressure p of the gas and on the collision cross section σ_c. The total velocity v is the vector sum of drift velocity v_D and thermal velocity.

The positive charge carriers drift to the electrode K and the negative ones to the anode A.

On their path from the position of creation to the electrodes the charge carriers can recombine. The number of recombination events depends on the time between creation and arrival at the electrodes and therefore decreases with increasing field strength E. As long as the number $Z = I/q$ of charge carriers that arrive at the electrodes per unit time is small compared to the recombination rate, the equilibrium between creation and recombination is not significantly disturbed. From (2.24) we get for the current density j at the electrodes according to (2.3) and (2.6b) and with the mobilities $u^\pm = \sigma_{el}^\pm/(n \cdot q)$

$$\begin{aligned} \boldsymbol{j} &= q \cdot n_{stat}(u^+ + u^-) \cdot \boldsymbol{E} \\ &= e\sqrt{\alpha/\beta}(u^+ + u^-) \cdot \boldsymbol{E}, \end{aligned} \tag{2.27}$$

if each charge carrier holds the elementary charge $\pm e$. This illustrates that Ohm's law (2.6b) is valid in this range. The current $I = j \cdot A$ at the electrodes with surfaces A and separated by the distance d increases linearly with the voltage $U = E \cdot d$.

If the voltage increases further the rate of recombination decreases because the drift velocity v_D increases and therefore the time during which charge carriers stay in the plasma where recombination takes place, decreases. Saturation of the current $I(U)$ is reached if all generated charge carriers reach the electrodes before they can recombine. For a creation rate α per unit volume and a distance d between the electrodes, the creation rate of pairs of charge carriers in the volume $V = d \cdot A$ is $n = \alpha \cdot d \cdot A$ and the saturation current density $j_{sat} = I/A$ is

$$j_{sat} = 2\alpha \cdot e \cdot d, \tag{2.28}$$

The factor 2 takes into account that positive as well as negative charge carriers contribute to the total current.

Example

1. The cosmic radiation creates in the layers of our atmosphere near the surface of the earth about $10^6\ m^{-3}\ s^{-1}$ ion pairs. The coefficient of recombination at 1 atm is about $\beta = 10^{-12}\ m^3\ s^{-1}$. With (2.24) we then get a stationary ion pair concentration of $10^9\ m^{-3}$.

The mobility u of positive ions in air at atmospheric pressure ($n_{\text{neutral}} \approx 3 \times 10^{25}\,\text{m}^{-3}$) and a collision cross section $\sigma_{\text{St}} \approx 10^{-18}\,\text{m}^2$ is

$$u = \frac{e}{m \cdot \bar{v} \cdot n \cdot \sigma_{\text{St}}} = 3 \times 10^{-4}\,\text{m}^2/\text{Vs},$$

The mean mobility of negative ions and electrons is higher by about the factor two.
Applying a voltage $U(V)$ at a parallel plate capacitor in air with its plate distance d (m) an electric current I flows because of the ion concentration in air with the current density according to (2.27)

$$j = e \cdot \sqrt{\alpha/\beta} \cdot (u^+ + u^-) \cdot E \Rightarrow$$
$$j = 1.5 \times 10^{-13} U/d, \quad [j] = \text{A/m}^2$$

In case of saturation all the created charge carriers reach the electrodes. The current density is then

$$j_{\text{sat}} = 2 \times 10^6 \cdot 1.6 \times 10^{-19} \cdot d.$$

For $d = 0.1\text{m}$ is $j_{\text{sat}} = 3.2 \times 10^{-14}\,\text{A/m}^2$. In this case saturation is reached already at a field strength E = 0.2 V/m.

2. Increasing the creation rate for example by irradiation with X-rays, up to $\alpha = 10^{12}\,\text{m}^{-3}\,\text{s}^{-1}$ ion pairs, the saturation field strength increases by a factor 10^3 to 200 V/m for the same recombination coefficient β.

If the voltage U across the electrodes is raised above the critical value U_C the charge carriers get enough energy by the electric field to ionize the neutral atoms or molecules of the gas by collisions (ionization by collision). The main contribution comes from the electrons because they have the same mass as the electrons of the neutral atoms and therefore the transfer of energy is more efficient (see Vol. 1, Chap. 4).

2.7.4 Mechanism of Gas Discharges

New charge carriers can be created by collision ionization only, if the electrons gain between two successive collisions sufficient energy in the electric field to ionize the neutral particles. An electron moving in x-direction over the mean free path Λ_x gains the energy $e \cdot E \cdot \Lambda_x$ in an electric field $\boldsymbol{E}$ in x-direction. The condition for ionization by collision over the distance Λ_x is therefore

$$e \cdot E \cdot \Lambda_x \geq W_{\text{ion}}. \tag{2.29}$$

A current of N electrons per unit time that are accelerated in x-direction by the field E creates along the distance dx

$$\text{d}N = \gamma N \text{d}x \tag{2.30}$$

new charge carrier pairs and therefore dN additional electrons that can ionize by collision after having been accelerated (Fig. 2.39). The factor

$$\gamma = \frac{(\text{d}N/N)}{\text{d}x}$$

gives the mean number of secondary electrons that are created by one primary electron on its way dx = 1 m along the x-axis. The mean free path $\Lambda \propto 1/p$ depends on the pressure inside the discharge volume. Therefore also the ionization capability γ depends on the ratio of E/p of field strength E and pressure p, and also on the ionization energy W_{ion}. Figure 2.41a shows $\gamma(E/p)$ for various gases. We see for example that at equal values of E/d the ionization capability of Ne or He is considerably lower than that of air. The reason is the higher ionization energy of Ne and He.

Integrating (2.30) yields the number of electrons after the distance $x = d$

$$N_1 = N_0 \text{e}^{\gamma d}, \tag{2.31}$$

where $N_0 = N(x = 0)$ is the number of electrons at $x = 0$ created, for example, by thermal emission from the cathode. The number $N^+ = N_0(\text{e}^{\gamma d} - 1)$ of positive ions created by collisional ionization per unit time are accelerated in the direction of the field and hit the cathode where they release secondary electrons. Here the N_0 primary electrons contained in (2.31) have to be subtracted). If δ is the mean number of secondary electrons created per ion the total number of secondary electrons is $\delta \cdot N_0(\text{e}^{\gamma d} - 1)$. The coefficient δ depends on the material of the cathode as well as on the kind of ions and their energy. The secondary electrons are accelerated towards the anode and create on this path length d

$$N_2 = \delta \cdot N_0 \cdot (\text{e}^{\gamma d} - 1) \cdot \text{e}^{\gamma d}$$

ion pairs. This process continues and we get a total of

$$N = N_0 \text{e}^{\gamma d} \sum_i \delta^i (\text{e}^{\gamma d} - 1)^i \tag{2.32}$$

secondary electrons per unit time (Fig. 2.40). For $\delta \cdot (\text{e}^{\gamma d} - 1) < 1$ the geometric series (2.32) gives

$$N = N_0 \frac{e^{\gamma d}}{1 - \delta(e^{\gamma d} - 1)}. \tag{2.33a}$$

The discharge current

$$I = eN = eN_0 \frac{\text{e}^{\gamma d}}{1 - \delta(\text{e}^{\gamma d} - 1)} \tag{2.33b}$$

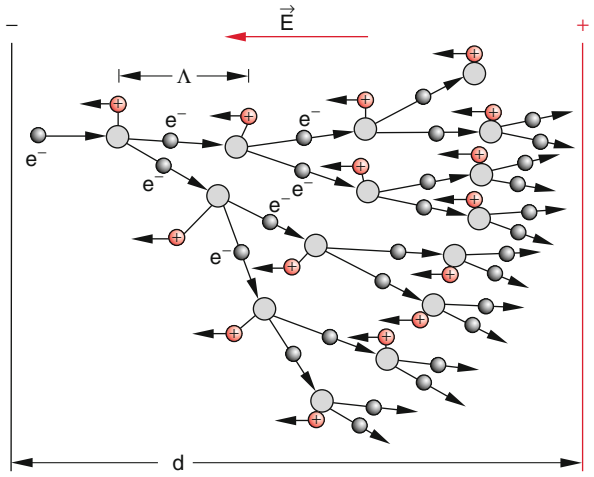


Fig. 2.40 Avalanche of electron-ion pairs created by collisional ionization in a gas discharge

increases more than linearly with the field strength E because γ and thus N increase steeply with E (Fig. 2.41a). As long as $\delta \cdot (e^{\gamma d} - 1) < 1$ the discharge is non-self-maintained. The current (2.33b) becomes zero if the number of primary electrons at $x = 0$ is zero ($N_0 = 0$) and no charge carriers are generated from outside, for example by X-rays or by a glow cathode.

This changes, if the ionization capability γ becomes so large that $\delta(e^{\gamma d} - 1) \geq 1$ or

$$\gamma \geq \frac{1}{d} \ln\left(\frac{\delta + 1}{\delta}\right), \tag{2.34}$$

because then the number N of charge carriers in (2.32) becomes infinite ($N \to \infty$).

For $N \to \infty$ the discharge becomes self-sustaining. For each accidentally created primary electron (created for example by cosmic radiation) an infinitely increasing avalanche of charge carriers is created. Since the ionization capability γ increases steeply with the field strength, the ignition condition (2.34) is fulfilled for every discharge above the ignition field strength E_Z. The discharge is self-sustaining. The ignition voltage U_Z depends on the kind of gas and the gas pressure (Fig. 2.41b) and on the geometry of the gas container, the distance d of the electrodes, their form, and of their material (because δ depends on the material of the cathode).

The condition for a self-sustaining discharge can be formulated as

Each charge carrier has to provide its own substitute.

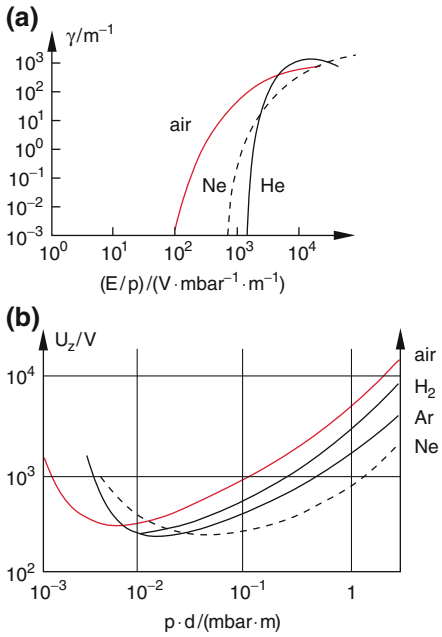


Fig. 2.41 **a**) Ionization capability γ as a function of the ratio E/p **b**) Ignition voltage of a gas discharge as a function of the product p times d of pressure p and distance d between the electrodes

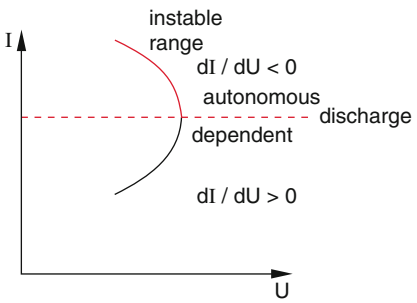


Fig. 2.42 Current-voltage characteristics of a gas discharge around the transition between stable and unstable discharge conditions

Note Since the conductivity σ_{el} increases with increasing density n of charge carriers, the resistance of the self-sustaining gas discharge decreases with increasing current (Fig. 2.42) and the current-voltage characteristic dI/dU becomes negative! Therefore the discharge current can increase unlimited at a constant voltage U. This would destroy the power supply or blow the fuse. To avoid such failures an Ohmic series resistance is included in the circuit to limit the current (Fig. 2.43).

In case of alternating current discharges a coil of inductance L and resistance $R = \omega \cdot L$ (see Sect. 5.4) is a better choice to stabilize the discharge.

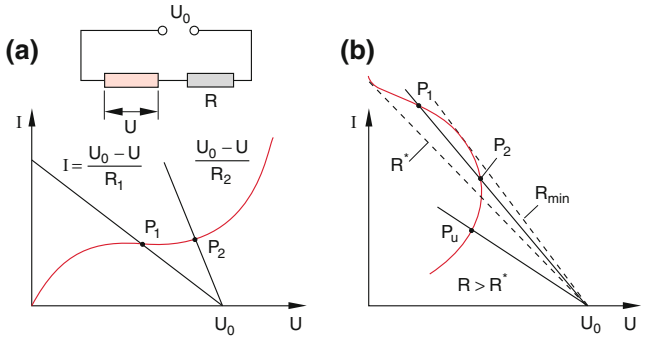


Fig. 2.43 Stable operation points P of a gas discharge with a resistor in series with the discharge, shown for different resistors

The increasing current generates an increasing voltage drop $\Delta U = I \cdot R$ at the resistance and only the fraction

$$U = U_0 - R \cdot I$$

remains at the discharge. A stable operation of the self-sustaining discharge adjusts itself at the intersection of the straight line of resistance $I = (U_0 - U)/R$ and the current-voltage characteristic $I(U)$ of the gas discharge (Fig. 2.43).

2.7.5 Various Types of Gas Discharges

During the collision with atoms the electrons can not only ionize but also transfer energy $W < W_{ion}$ that can excite the neutral atoms. In general these excited states release their energy W within a short time (typically 10^{-8} s) by emitting light with the photon energy $W = h \cdot \nu$. That's why gas discharges glow. Also during the recombination of electrons and ions light is emitted. Intensity, color and spatial distribution of the light emission depends on the gas itself, its pressure, and on the type of gas discharge. We distinguish the following types.

2.7.5.1 Glow Discharge

Glow discharges are discharges at low pressure ($p = 10^{-4}$–10^{-2} bar) and low currents of a few mA. We see shining layered structures (Fig. 2.44), which change with pressure p and discharge voltage U. The observed structure corresponds to the distribution of the electric field strength $E(x)$ that is no more constant (Fig. 2.45).

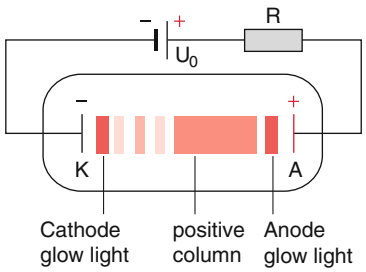


Fig. 2.44 The different regions of a glow discharge

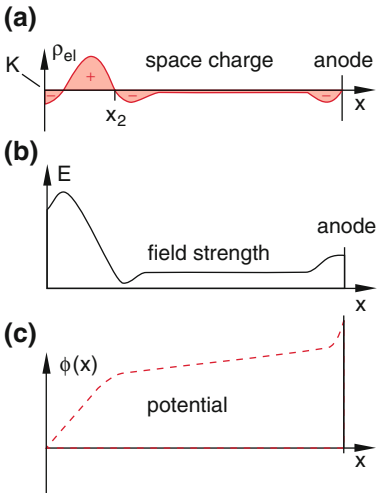


Fig. 2.45 Space chare distribution, electric field strength and potential along the axis of a glow discharge

The secondary electrons created by the incident ions at the cathode are accelerated until they have gained along the distance x enough energy to excite the gas atoms. Therefore next to the cathode the *negative glow light* is created. After passing the distance x_2 the electrons have enough energy to ionize. At this position a great number of electron-ion pairs is generated. The heavier ions leave the small volume slower in the direction to the cathode than the electrons which leave to the anode and the result is a surplus of positive charges. Therefore the volume charges at x_2 increase the field strength between cathode and position x_2: This is called the cathode fall of the potential in Fig. 2.45c. Thus the field strength between position x_2 and anode is reduced correspondingly. In this area the acceleration of the electrons is reduced and consequently the rate of ionization. This space is filled with negative volume charges (Fig. 2.45a).

The largest part of the discharge volume is filled with the positive column with a relatively constant electric field, which is strong enough to maintain an equilibrium between ionization and recombination. Here the electrons have enough energy to excite the atoms and the entire positive column radiates diffuse light.

At decreasing pressure the mean free path becomes larger and the positive column divides into many layers with a mutual distance Λ_x corresponding to the mean free path.

2.7.5.2 Arc Discharge

Now we will discuss discharges of high currents and higher pressure of the surrounding gas. Because of the high currents the electrodes become very hot and can supply electrons by thermal emission. Therefore the further supply of electrons does no longer need the impact of ions. The electric conductivity of the arc discharge is very high so that after the ignition the voltage across the arc decreases and the arc continues

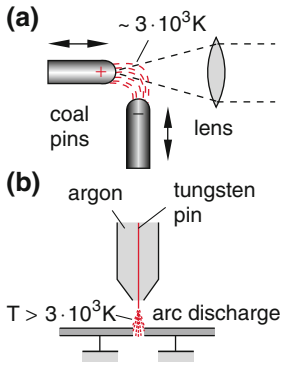


Fig. 2.46 **a**) Arc discharge between two carbon electrodes and imaging of the bright light for use in brilliant projection light sources **b**) arc welding of metals with argon as shield gas

already at low voltages. An example is the carbon arc lamp (Fig. 2.46a). It is used as a bright source of light for projections. Also electrical welding uses such high current arcs between the workpiece and a rod of tungsten as the two electrodes (Fig. 2.46b). To ignite the arc both electrodes are brought into contact for a short time. To avoid corrosion of the workpiece an inert shielding gas such as argon is blown around the electrodes.

Also mercury or xenon high pressure arc lamps of high brilliant intensity are examples of high current arc discharges.

They are ignited by a short high voltage pulse and then they shine as a self-sustaining discharge

2.7.5.3 Spark Discharge

Spark discharges are short-time arc discharges that extinguish because the voltage across the distance of the arc breaks down. A typical example is the discharging of a high voltage capacitor through a gas discharge tube. A practical application is the use of flash lights in photography to illuminate or to brighten objects. A spectacular form of discharges can be observed as lightning in thunderstorms. Very short but heavy heating up of the air in the discharge channel creates a volume of high pressure that propagates as an acoustic shock wave (thunder) through the air.

More extensive descriptions of gas discharges can be found in the literature [8–10].

2.8 Current Sources

Up to now we have dealt with the mechanism of transporting electric charges through solids, liquids, and gaseous conductors, but we have not discussed the generation of electric currents.

All current sources are based on the separation of positive and negative charges.

During this spatial separation work has to be done against the attractive Coulomb forces. This work comes from mechanical or chemical energy, from light or nuclear energy. The separation of charges causes a potential difference between the spatially separated charges inside the current source. This can be measured as a voltage U across its terminals or poles. Connecting these poles by a conductor enables the flow of a current I. Its maximum value is smaller than that defined by the open circuit voltage U and the Ohmic resistance R of the external connection $I_{max} < U/R$. This is due to the limited internal production of charges $I = dQ/dt$ and can be represented by a resistance called **internal resistance** that adds to the external resistance R.

By far the most often used technical method to produce currents are the electrodynamic generators. They use magnetic induction to separate the charges. We will treat them in Chap. 5.

An important role in creating current sources which are independent of the public power network play chemical current supplies as batteries or accumulators. Especially advanced fuel cells, which are the subject of intense research, will become more important for electrically driven cars. We will explain both types of chemical current supplies briefly.

The principle of solar cells where the light of the sun is used to generate electric power (Fig. 2.47) cannot be explained until Vol. 3 where we treat semiconductors.

Finally we will present thermo-electricity which uses the dependence of the contact voltage between different metals on temperature.

Fig. 2.47 Solar cell field as competion to coal power stations in the background

2.8.1 Internal Resistance of Current Sources

Each current source has an internal resistance R_i because the charge carriers suffer collisions with the atoms or molecules of the corresponding conductor on their path from the position of the charge separation to the output terminals. When an external resistor R_e is connected between the terminals the open circuit voltage U_0 of the current supply (this is the voltage without load, also called the emf = electro-motive force) decreases to the lower value

$$U = U_0 - I \cdot R_i = U_0 \cdot \left(1 - \frac{R_i}{R_i + R_e}\right) = U_0 \frac{R_e}{R_i + R_e}. \tag{2.35}$$

where $I = U_0/(R_i + R_e)$ is the current through the load (Fig. 2.48).

The voltage across the load becomes dependent on the load!

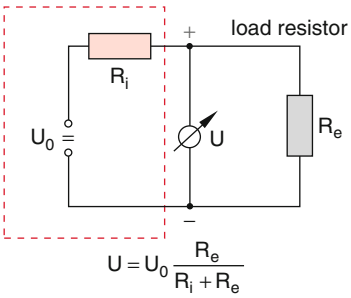


Fig. 2.48 Equivalent circuit diagram of a current source with internal resistor R_i and external load resistor R_e

Note The internal resistance R_i can be made very small by an electronic stabilizer so that the external voltage is nearly constant within a given range of the current I.

2.8.2 Galvanic Cells

If we immerse two different metals into a liquid electrolyte an electric voltage is generated between the two electrodes. The reason for this voltage can be explained as follows.

Between the metal electrode and the surrounding electrolytic liquid is a concentration gradient of metal ions which is counterbalanced by diffusion, i.e. by transfer of metal ions from the electrodes into the liquid. However, the binding-energy $|e \cdot \phi_1|$ of the metal ions in the metal

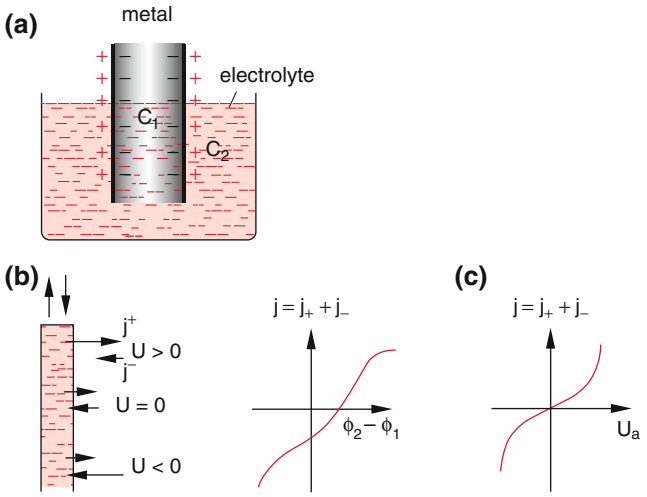


Fig. 2.49 **a**) Metal electrode immersed into an electrolyte **b**) potential difference between electrode and liquid for different current densities of ion currents leaving or penetrating the electrode **c**) total current density as a function of the potential difference and of an external voltage

electrode is generally much larger than the binding-energy $|e \cdot \phi_2|$ of the ions in the liquid, which is determined by the accumulation of the ions to the water dipole molecules. Therefore only a small part of the metallic ions dissolve into the liquid where they form a narrow layer of positive charges around the electrodes, whereas the electrodes become negatively charged because of the missing positive ions (Fig. 2.49). This results in a potential difference $\Delta\phi$ between electrode and electrolyte which drives the ions back into the metal. Equilibrium is reached if the rate of dissolving ions equals that of the ions returning into the metal.

The ratio of the concentrations c_1 in the metal and c_2 in the liquid at equilibrium is given by the Boltzmann distribution

$$\frac{c_1}{c_2} = e^{-U/kT} \tag{2.36}$$

(see the equivalent discussion of the barometric formula in Vol. 1, Chap. 7).

Under this equilibrium condition no current flows. If an external positive voltage U_a is applied between electrode and electrolyte more positive metal ions pass into the solution until the electrode is completely dissolved. In case of a negative external voltage the potential of the electrode is lowered and more positive ions can accumulate at the electrode.

For the Galvanic cell with two different metal electrodes (Fig. 2.50) the binding energy of the metal ions differs and therefore also the potential differences $\Delta\phi_i$ between the electrolyte and the two electrodes differ. This results in a voltage

$$U = \Delta\phi_1 - \Delta\phi_2$$

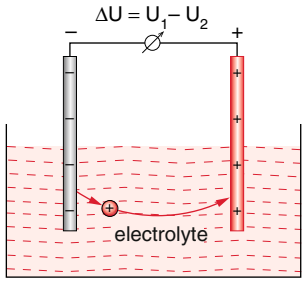


Fig. 2.50 Two electrodes consisting of different materials immersed into a liquid electrolyte form a galvanic cell

Table 2.6 Electrochemical series of some metals measured at $T = 293$ K against the reference of the hydrogen normal electrode for a concentration of 1 mol ions per 1 liter electrolytic aqueous solution

Electrode	*U*/V	Electrode	*U*/V
Li^+/Li	−3.02	Ni^{++}/Ni	−0.25
K^+/K	–2.92	Pb^{++}/Pb	−0.126
Na^+/Na	–2.71	$H_2/2H^+$	0
Zn^{++}/Zn	–0.76	Cu^+/Cu	+0.35
Fe^{++}/Fe	–0.44	Ag^{++}/Ag	+0.8
Cd^{++}/Cd	–0.40	Au^{3+}/Au	+1.5

between the electrodes. The metals can be arranged in an electro-chemical series (Table 2.6) with increasing standard electrode potential which is defined as the potential difference against a standard hydrogen electrode under standard conditions ($T = 298.15$ K, concentration of 1 mol/l of the electrolyte).

Note The electro-chemical series is not identical to the contact potential series of Table 1.3, where the work functions for electrons are listed instead of the binding energy of the ions.

One example of a Galvanic cell is the copper-zinc galvanic cell (Fig. 2.51). A Zinc electrode is immersed into a $ZnSO_4$ solution and the copper electrode into a $CuSO_4$-solution. Since the binding energy of the Zn^{++} ions is smaller than that of the copper ions, more Zn-ions are transferred into the solution and leaving a surplus of electrons. Connecting the two electrodes by a conducting wire the electrons flow to the copper electrode which forms the positive pole of the Galvanic cell whereas the Zn-electrode forms the negative pole.

Instead of the two half-cells in Fig. 2.51a both electrodes can be immersed into the same electrolytic solution (Fig. 2.51b). The voltage delivered by a Galvanic cell is equal to the potential difference between the two metals and can be immediately obtained from the electro-chemical series (Table 2.6) For example in a Zn-Cu Galvanic cell with a Zn and a Cu-electrode in a dilute H_2SO_4 solution the voltage between the electrodes is 1.1 V where the Zn-electrode forms the negative pole and the Cu-electrode the positive pole.

Connecting the two electrodes by an external load resistance R_e one measures the current

$$I = U/R_e + R_i,$$

where R_i is the internal resistance of the Galvanic cell and the current is carried by electrons which flow from the negative Zn-electrode to the positive Cu electrode.

Luigi Galvani (Fig. 2.52) was the first scientist who discovered in 1780 that a voltage appeared if he connected

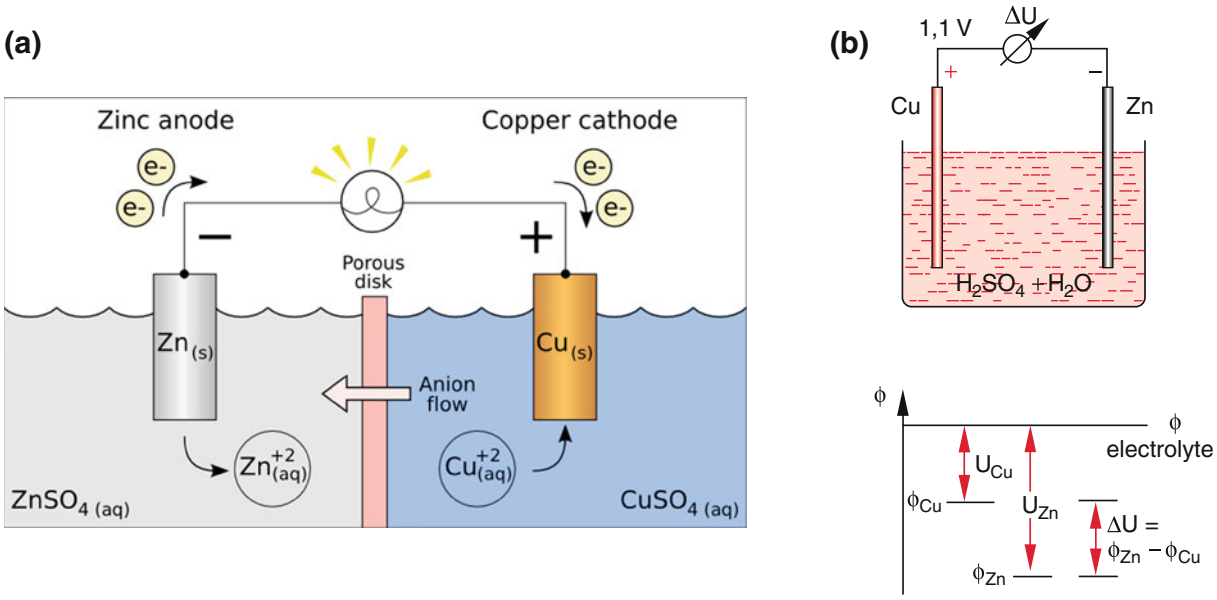


Fig. 2.51 **a**) Cu-Zn Galvanic Cell with two separated parts (The original uploader was Ohiostandard at English Wikipedia. [CC BY-SA 3.0]) **b**) Galvanic cell with the different potential differences between electrodes and electrolyte

Fig. 2.52 Luigi Galvani

Fig. 2.53 Alessandro Volta

two different metals with a frog's leg. The unit of the voltage 1 V is named after Alessandro Volta (1745–1824 Fig. 2.53). In Fig. 2.54 the different currents of ions and electrons in a Galvanic cell are illustrated.

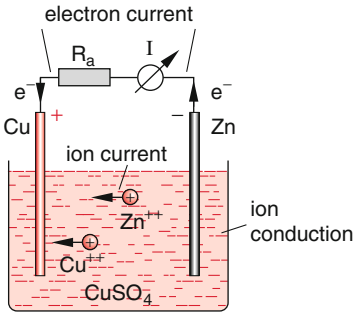


Fig. 2.54 The currents within a galvanic cell with external load

2.8.3 Accumulators

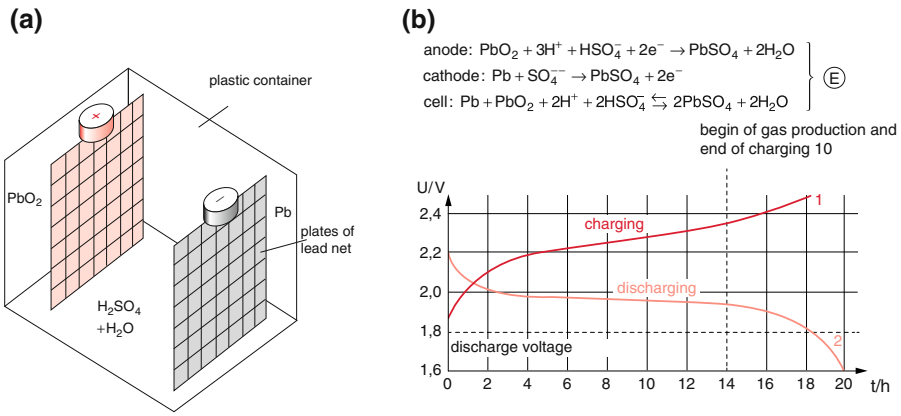
If two lead plates are immersed into a sulfur acid solution diluted with water both plates are soon covered by a layer of lead sulfate $PbSO_4$. When now an external voltage is applied to the plates (Fig. 2.55a) the ions H^+ and OH^-, which have been dissociated in the electrolyte (see Sect. 2.6) move to the electrodes. Here the ions deliver their charge and react with the $PbSO_4$ layers according to the following scheme:

$$\begin{aligned} \text{Anode: } & PbSO_4 + 2OH^- \rightarrow PbO_2 + H_2SO_4 + 2e^-; \\ \text{Cathode: } & PbSO_4 + 2H^+ + 2e^- \rightarrow Pb + H_2SO_4. \end{aligned} \tag{2.38}$$

During this charging process the anode converts to lead oxide PbO_2 and the cathode to metallic lead Pb. The charging process has created a Galvanic cell with two unequal electrodes, which now can deliver a voltage between the two poles. Between the plus-pole (PbO_2) and the minus pole (Pb) appears a voltage of 2 V.

At the end of the charging process one observes the production of oxygen gas at the anode (from the

Fig. 2.55 Lead accumulator **a**) schematic design **b**) charging and discharging curves *U(t)*



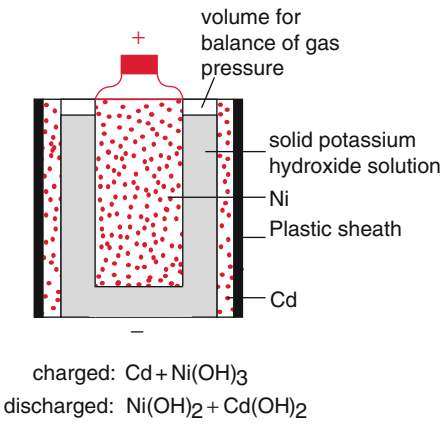


Fig. 2.56 Nickel-Cadmium battery

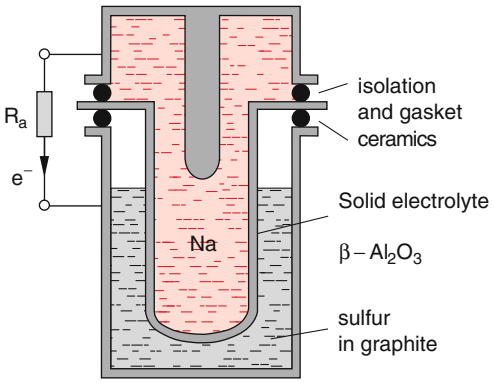


Fig. 2.57 Solid sodium-sulphur battery

reaction $4OH^- \rightarrow 2H_2O + O_2 + 4e^-$) and of hydrogen gas at the cathode ($2H^+ + 2e^- \rightarrow H_2$).

When discharging the accumulator the processes (2.38) proceed into the opposite direction:

$$\begin{aligned} &\text{Anode:} \quad PbO_2 + HSO_4^- + 3H^+ + 2e^- \rightarrow PbSO_4 + 2H_2O \\ &\text{Cathode:} \quad Pb + SO_4^{--} \rightarrow PbSO_4 + 2e^-. \end{aligned} \tag{2.39}$$

The temporal variation of the output voltage during these process is shown in Fig. 2.55b). The efficiency η of the accumulator is defined as the ratio of delivered energy during the discharge to the energy supplied for charging. It amounts to about 75–80%. The residual 20–25% are wasted into heat.

The storage capacity is about 30 Wh per kg lead. In order to enlarge the surface of the electrodes Pb-grids are used. Technical details about Pb-accumulators can be found in [10–12].

2.8.4 Different Types of Batteries

Besides the lead accumulator discussed in the foregoing section, there are several other electric current sources which are based on charge separation by chemical reactions. One example is the rechargeable Nickel-Cadmium battery (Fig. 2.56). Here Ni and Cd-electrodes are immersed into a KOH-solution, which are covered by a hydroxide layer. During the charging process the reactions

$$\begin{aligned} &\text{Cathode:} \quad Cd(OH)_2 + 2e^- \rightarrow Cd + 2OH^- \\ &\text{Anode:} \quad Ni(OH)_2 + 2OH^- \rightarrow 2Ni(OH)_3 + 2e^- \end{aligned}$$

are initiated. The charging process ends when the whole surface of the cathode has been converted to cadmium. During the discharging process the reactions are inverted in time and in the outer part of the circuit the electrons flow from the cadmium to the nickel electrode. The whole battery is enclosed in a gas-tight plastic mantle.

Remark

Since 2006 the Ni-Cd-batteries are forbidden because of the toxic cadmium.

For many applications the ratio of stored energy and weight of the lead accumulator is too bad. Furthermore the liquid acid is often not acceptable. One therefore has looked for solid electrolytes. One solution is the sodium-sulfur battery with an electrolyte which consists of solid Al_2O_3 ceramics (Fig. 2.57). On one side of the electrolyte is liquid sodium, on the other side liquid sulfur which is absorbed by a graphite sponge, in order to increase the electric conductivity. At the anode the reaction

$$2Na^+ + S + 2e^- \rightarrow Na_2S$$

At the cathode the reaction

$$Na \rightarrow Na^+ + e^-.$$

proceeds. The output voltage is about 2 V and the maximum energy density is with 1 kWh/kg by a factor of 30 higher than in the Pb-accumulator.

For small independent electronic devices, such as radios, mobile telephones or tape recorders small “dry batteries” as modern devices of the old Leclanchè-Elements (Fig. 2.58) are used. A graphite rod doped with MnO_2 is the central positive electrode, whereas the outer zinc cylinder forms the negative pole. The solid electrolyte consists of a NH_4Cl solution fixed in cellulose between the two electrodes.

A particular efficient and rechargeable device is the **lithium-ion-accumulator**. Its energy storage capacity is with 200–500 Wh/kg much higher than that of the Pb-accumulator. Therefore it is nowadays used nearly exclusively for laptops. i-phones and mobiles. Its basic principle is explained in Fig. 2.59.

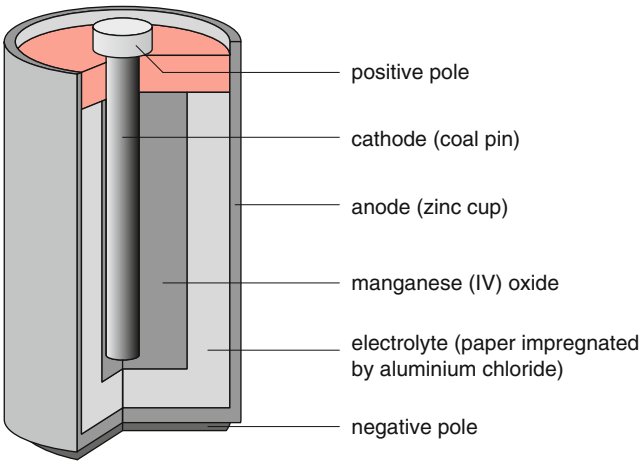


Fig. 2.58 Solid magnesium-zink battery

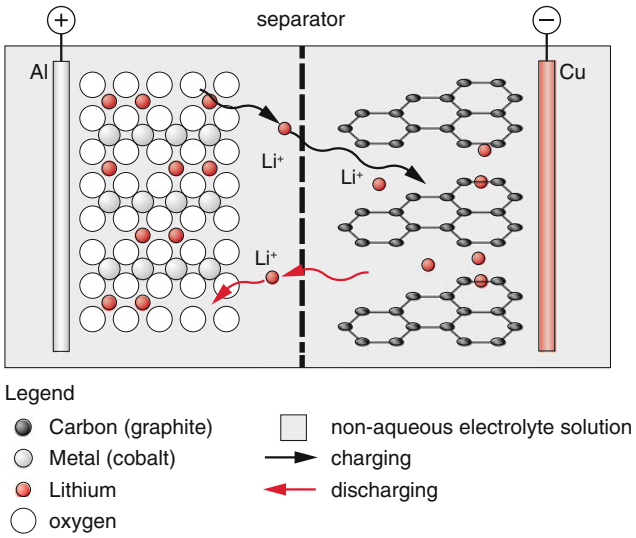


Fig. 2.59 Lithium-ion-accumulator

The anode consists of a metal-oxide compound Li-M-oxide (M = transition element e.g. Ni, Fe, Co, etc.), whereas the cathode material is a Li-Graphite compound. The electrolyte between the electrodes consists of a water-free gelatinous solid (e.g. Lithium-tetrafluorborat $LiBF_4$) which contains on the anode side a lattice structure of Oxygen-Cobalt and Lithium atoms, on the cathode side graphite compounds. The two sides are separated by a micro-porous membrane (separator), which is transparent for the Li^+-ions.

During the discharging process the cathode delivers electrons to the consumer circuit, which drift to the anode through this external connection. The cathode then becomes more positive and the anode more negative, which allows the Li^+-ions to pass through the separator to the anode. Since as many Li^+-ions migrate from the cathode to the anode as

Fig. 2.60 Lithium-ion button cell

electrons the electrodes remain neutral, i.e. the potential of the electrodes stays nearly constant. During the charging process the direction of ion-migration is reversed.

During charging the negative electrode acts a s cathode, during the discharging as anode, whereas the positive electrode acts during charging as anode and during discharging as cathode. The electrons can move freely inside the electrodes while the ions move free in the electrolyte.

The reactions during the charging and discharging process can be described by the following equations:

Discharging:

Negative electrode:

$$\mathrm{Li}_x + \mathrm{C}_n \rightarrow n \cdot \mathrm{C} + x \cdot \mathrm{Li}^+ + x \cdot \mathrm{e}^-$$

Positive electrode:

$$\mathrm{Li}_{1-x}\mathrm{Mn}_2\mathrm{O}_4 + \mathrm{Li}_x\mathrm{C}_n \rightarrow \mathrm{LiMn}_2\mathrm{O}_4 + C_{n'}$$

Charging process:

$$\mathrm{Li}_{1-x}\mathrm{Mn}_2\mathrm{O}_4 + \mathrm{Li}_x\mathrm{C}_n \rightarrow \mathrm{LiMn}_2\mathrm{O}_4 + C_n$$

The $LiCoO_2$ accumulator delivers an output voltage of 3.6 V, which is about three times of the output voltage of the Ni-metal-hydride accumulator.

Lithium-ion batteries are available also as small button cells (Fig. 2.60) for electronic devices that do not consume much energy. The diameter of the cell is about 1 cm, for hearing devices about 0.7 cm.

More information about rechargeable batteries can be found in [12, 13].

2.8.5 Fuel Cells

In the accumulator the chemical energy of the reaction partners, (PbS + H_2O) resp. (Pb + H_2SO_4) contained in the accumulator is transformed into electrical energy. The reaction products remain inside the cell and lead to the decrease of the output voltage (discharging). The energy storage capacity of batteries and accumulators is therefore limited.

This disadvantage is avoided in chemical fuel cells, because here the reaction partners are continuously supplied from outside. In Fig. 2.61 a simplified scheme of a fuel cell operating with hydrogen and oxygen, is illustrated. Here electric energy is generated by the exothermic oxy-hydrogen reaction

$$2H_2 + O_2 \rightarrow 2H_2O. \tag{2.40}$$

which proceeds in the fuel cell under controlled conditions, in order to avoid an explosion-like energy release. The trick of the fuel cell is the spatial separation of oxidation- and reduction reactions. The reaction (2.40) is split by a suitable construction of the fuel cell into the part

$$O_2 + 2H_2O + 4e^- \rightarrow 4OH^- \tag{2.40a}$$

proceeding at the cathode which delivers one electron per OH-radical (electron acceptance = reduction) and the part

$$2H_2 + 4OH^- \rightarrow 4H_4O + 4e^-, \tag{2.40b}$$

at the anode where one electron is delivered per OH-radical (oxidation reaction).

For both reactions a catalyst as well as an electrolytic solution in water are necessary. Therefore the reaction can only take place at the boundary between gas, electrolyte and catalyst. This demands a special form and arrangement of the electrodes. One uses for instance porous electrodes into which the supplied gas (O_2 resp. H_2) as well as the electrolyte can penetrate. The three-phase-boundary corresponds to the meniscus of the electrolyte (Fig. 2.61b) in the pores of the electrode, which is realized at equilibrium between gas pressure and liquid capillary pressure (Vol. 1, Sect. 6.4). In order to reach this equilibrium the diameter of the pores has to be of the correct size. As catalysators for the cathode (H_2-elecrode) Nickel is used, while for the anode (O_2 electrode) silver is a good choice.

In the reactions (2.40a, b) two water molecules (H_2O) are formed from $2H_2$ molecules and $1O_2$ molecule. These reactions are exothermic and the energy of about 5 eV is delivered, because the binding-energy of the two water molecules is $2 \cdot 9.5 = 19$ eV, that of two H_2. Molecules $2 \cdot 4.5$ eV and that of one O_2 molecule is 5.1 eV. The excess energy, which is released is therefore $E_e = 19 - 9 - 5.1 = 4.9$ eV.

Typical output powers of such fuel cells are about 0.5 W per cm^2 electrode surface at a voltage of 0.8 V. Since this voltage is for many applications too low one has to connect several cells in series. In Fig. 2.62 output voltage and output power of 33 fuel cells connected in series are shown in dependence of the consumed current. Nowadays power densities of 0.2 kW per 1 kg cell weight can be realized. This is larger by one order of magnitude than that of Pb-accumulators.

The great advantage of such fuel cells is the direct conversion of chemical energy into electric energy, without the detour via thermal energy (which is necessary for thermal power stations). Therefore here the limitation set by the Carnot-efficiency (see Vol. 1, Chap. 10), is avoided. The main advantage is the avoidance of polluting exhaust gases, which represent an unsolved problem for combustion engines, because the exhaust of fuel cells is only harmless water vapor.

The main problem, unsolved up today is the membrane which should be strong enough to withstand the pressure of the supplied gases and the progressive poisoning of the catalyst by tiny concentrations of impurities in the supplied gases. Meanwhile it is, however, possible to develop powerful and long living fuel cells, which represent in combination with an electro-motor interesting alternatives to combustion engines for cars [14, 15]. While the combustion engines emit besides CO_2 gases also the toxic gases CO, NO and unburned hydro-carbon gases the exhaust of fuel cells consists only of harmless water vapor.

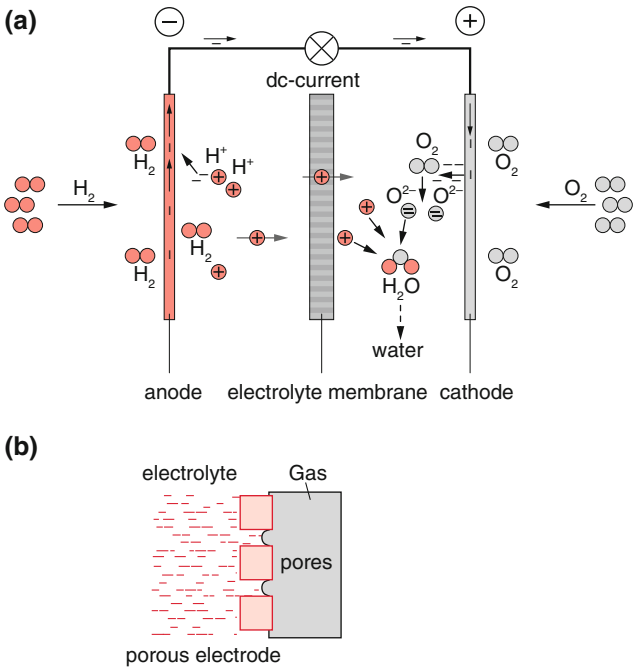


Fig. 2.61 Hydrogen-Oxygen fuel cell **a**) schematic operation **b**) section of separating membrane

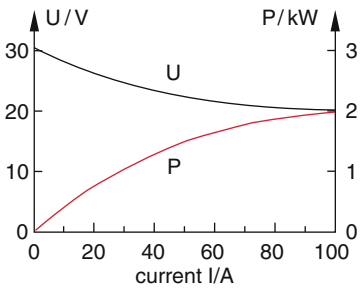


Fig. 2.62 Output voltage and power of 33 fuel cells connected in series as a function of the load current

Meanwhile already small fuel cells are commercial available as energy sources for flash lights, bicycle illumination etc. In some German cities public busses run already with fuel cells.

Several car manufacturers announced for 2019 to sell cars with fuel cells as sole drive source.

More information about chemical energy sources can be found in [16].

2.9 Thermal Current Sources

The temperature dependence of the contact potential between two different metals as well as the thermo-diffusion of conduction electrons in metals can be used for the generation of thermal current sources.

2.9.1 Contact Potential

In order to remove the freely moving conduction electrons in a metal out of the metal one has to supply work against the attractive forces between the negatively charged electrons and the positively charged ions of the crystal lattice. This **work function** W_a is analog to the evaporation energy of an atom which leaves the liquid into the gaseous phase (see Vol. 1, Sect. 10.4.2). If we choose the vacuum potential $\phi_{vak} = 0$, the work function for a metal with the highest electron energy state E_C (also called the *chemical potential* see Vol. 1, Sect. 10.8) becomes $W_a = -E_C$. The work function is negative because one has to supply energy to remove the electrons out of the metal.

If two different metals with different work functions W_{a_1} and W_{a_2} are brought into contact, electrons flow from the metal with the lower work-function into the metal with the higher work function, because this is energetically favorable. This charge separation causes a space charge at the boundary between the two metals (Fig. 2.63) which results in an opposite electric field that drives the electrons back. Equilibrium is reached if the currents in the two opposite directions just cancel.

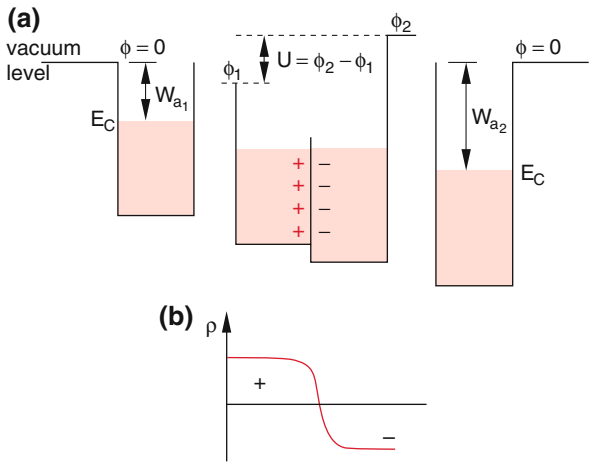


Fig. 2.63 **a**) Contact potential at the boundary between two metals with different work-functions **b**) Charge density around the contact of the two metals

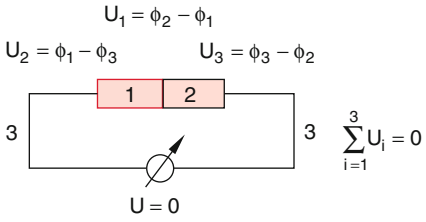


Fig. 2.64 The contact potential cannot be directly measured, because in a closed circuit the sum of all voltages is zero

These space charges shift the potentials ϕ of the two metals to ϕ_1 and ϕ_2 and a potential difference $U = \phi_2 - \phi_1$ develops which is called the **contact potential**.

However, this contact potential cannot be directly measured, because for the measurement a closed current loop must be realized (Fig. 2.64) where the sum of all contact potentials is zero.

2.9.2 Seebeck Effect

When two different electrical conductors A and B are connected to a circuit (Fig. 2.65) the voltmeter shows the voltage $U = 0$ as long as the two connections are kept at the same temperature (see foregoing section). If, however, the two contacts are at different temperature T_1 and T_2 the thermo-voltage

$$U = (S_A - S_B)(T_1 - T_2). \qquad (2.41a)$$

is measured. The coefficients S_A and S_B which depend on the two materials are called **Seebeck-Coefficients**. They are measured in units of [V/K]. Typical values for metals are 10^{-5}–10^{-6} V/K, while for semiconductors they are much larger, typically 10^{-3} V/K (Table 2.7). The Seebeck coefficients are temperature dependent (Fig. 2.66a) and for semiconductors they strongly depend on the concentration of impurity atoms (Fig. 2.66b). When the volt-meter in

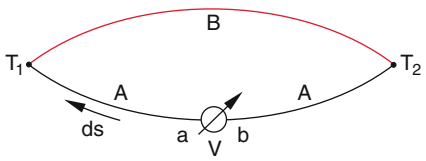


Fig. 2.65 Origin of thermoelectric voltage when the two contacts are kept at different temperatures

Table 2.7 Seebeck coefficients for some metals and semiconductors

Material	S [μV/K]
Mercury	0.6
Aluminum	3.5
Copper	6.5
Germanium	300
Tellurium	500
Selenium	900

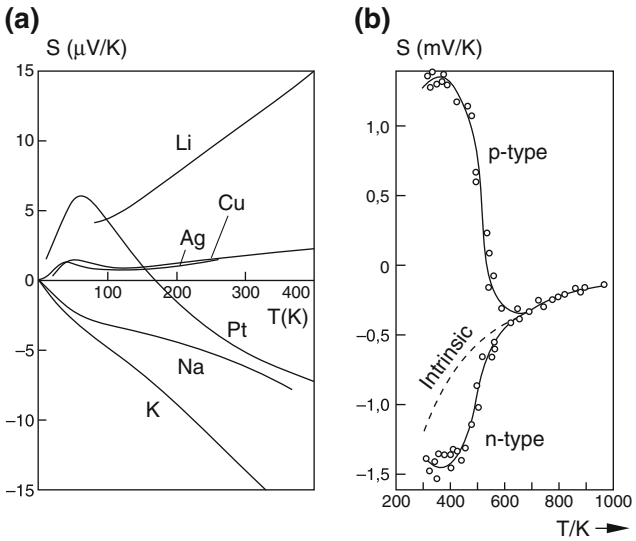


Fig. 2.66 Temperature dependence of Seebeck coefficients **a**) for some metals, **b**) for doped semiconductors for some metals

Fig. 2.65 is replaced by an ampere-meter a current $I = U/R$ is measured which depends on the thermo-voltage and the total resistance of the circuit.

The question is now: What is the cause of the thermo-electric-voltage [17]?

2.9.3 Thermoelectric Voltage

The contact voltage depends on the temperature of the contact. This can be explained as follows:

In Vol. 1, Sect. 7.3.5 it has been shown that under thermal equilibrium the concentrations n_1, n_1 of particles with different energies E_1 and E_2 follow the Boltzmann distribution

$$n_1/n_2 = \mathrm{e}^{-\Delta E/kT} \tag{2.41}$$

with $\Delta E = E_2 - E_1$.

Although the freely moving conduction electrons in metals do not generally follow a Boltzmann but a Fermi-distribution (see Vol. 3) we can approximate for $\Delta E \gg k \cdot T$ the electron distribution by (2.41). In this case the energy difference $\Delta E = -e \cdot (\phi_2 - \phi_1) = e \cdot U$ is given by the contact voltage U. Solving for U yields

$$U = \frac{k \cdot T}{e} \ln \frac{n_1}{n_2}. \tag{2.42}$$

If the two contacts in this closed circuit are at different temperatures the temperature dependent contact voltages

$$U_1 = \frac{kT_1}{e} \ln \frac{n_1}{n_2}, \quad U_2 = -\frac{kT_2}{e} \ln \frac{n_1}{n_2}$$

are different, The ratio n_1/n_2 of the electron concentrations is mainly determined by the different work functions of the two metals and only to a minor extent by the temperature (see Vol. 3).

Note The voltmeter in Fig. 2.64 does not measure the difference $\Delta U = U_1 - U_2$ between the points 1 and 2 but rather the total voltage between the points a and b (Fig. 2.67). This voltage can be composed of the potential differences

$$\begin{aligned} U = {} & [\phi_C(T_1) - \phi_B(T_1)] + [\phi_B(T_1) - \phi_B(T_2)] \\ & + [\phi_B(T_2) - \phi_A(T_2)] + [\phi_A(T_2) - \phi_A(T_1)] \\ & + [\phi_A(T_1) - \phi_C(T_1)] = 0. \end{aligned} \tag{2.42a}$$

This illustrates that, if the thermo-voltage would be solely caused by the different contact potentials the voltmeter would measure the voltage $U = 0$. Therefore there has to be another real cause. This is the temperature dependent thermo-diffusion of the electrons which results in a diffusion current with the density

$$j_{\mathrm{ThD}}(\boldsymbol{r}) = n \cdot \boldsymbol{u}(\boldsymbol{r}) \tag{2.42b}$$

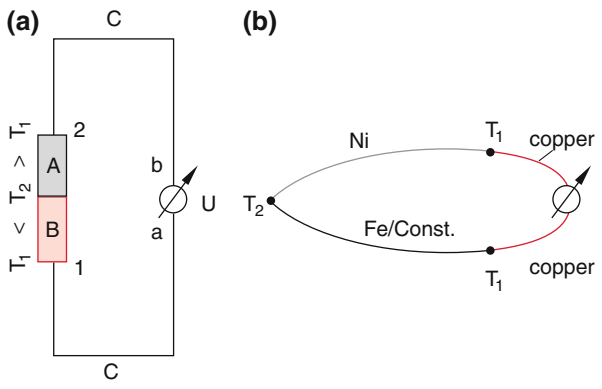


Fig. 2.67 Illustration of thermo-voltage **a**) circuit **b**) example of a thermo-element

where the thermo-diffusion velocity can be derived as follows [17, 18].

The conduction electrons in the conductor material collide with the lattice atoms (see Sect. 2.2.1). For the mean free path Λ of the electrons the velocity of the electron at the position $\boldsymbol{r}$ is determined by the temperature $T(\boldsymbol{r} - \lambda \boldsymbol{v}/|v|)$ at the location of the last collision. The mean velocity $\langle \boldsymbol{v} \rangle = v_D(r)$ is the drift velocity $v_D(r) = u(r)$. It can be obtained by averaging over all directions of the velocity:

$$\boldsymbol{u}(\boldsymbol{r}) = \langle \boldsymbol{v} \rangle_{\boldsymbol{r}} = \frac{1}{4\pi} \int \bar{v} \cdot \hat{v} \cdot T(\boldsymbol{r} - \lambda \hat{v}) \mathrm{d}\lambda \qquad (2.42c)$$

with the unit vector $\hat{v} = \boldsymbol{v}/|\boldsymbol{v}|$. The expansion of the integrand

$$T(\boldsymbol{r} - \lambda \hat{v}) \approx T(\boldsymbol{r}) - \lambda \hat{v} \cdot \nabla T(\boldsymbol{r})$$
$$\bar{v}(T(\boldsymbol{r} - \lambda \hat{v})) \approx \bar{v}(T(\boldsymbol{r})) - \lambda \hat{v} \nabla T(\boldsymbol{r}) \cdot \frac{\mathrm{d}\bar{v}}{\mathrm{d}T}$$

gives (see Problem 2.14):

$$\boldsymbol{u}(\boldsymbol{r}) = -\frac{\lambda}{3} \frac{\partial \bar{v}}{\partial T} \cdot \nabla T(r). \qquad (2.42d)$$

Besides the thermo-diffusion there is also the normal diffusion which depends on the concentration gradient and which is also existent for a spatially constant temperature (grad $T = 0$): The normal diffusion generates a particle current density

$$\boldsymbol{j}(\boldsymbol{r})_{\mathrm{Diff}} = -D \cdot \nabla n(\boldsymbol{r}). \qquad (2.42e)$$

Since the electrons carry the charge $-e$ their thermo-diffusion from a location at higher temperature to a location with lower temperature generates a spatial charge that causes an electric field $\boldsymbol{E}(\boldsymbol{r})$. This field generates in turn a drift of the charge carriers with the current density

$$\boldsymbol{j}(\boldsymbol{r})_{\mathrm{Drift}} = -\frac{\sigma}{\mathrm{e}} \boldsymbol{E}(r). \qquad (2.42f)$$

The total current density is then

$$\boldsymbol{j}_{\mathrm{total}} = \boldsymbol{j}_{\mathrm{Diff}} + \boldsymbol{j}_{\mathrm{ThD}} + \boldsymbol{j}_{\mathrm{Drift}}. \qquad (2.42g)$$

When the circuit in Fig. 2.67 is closed, a current with the current density $\boldsymbol{j}_{\mathrm{total}}$ flows through the circuit, which can be measured with the ampere-meter.

In the open circuit the spatial charge increases until the resulting drift current just compensates the other shares in (2.42g) and the total current density becomes zero. In this case the thermo-voltage appears at the voltmeter.

Note As mentioned before this thermo-voltage is due to the sum of all effects and would be zero if only the contact potentials would contribute.

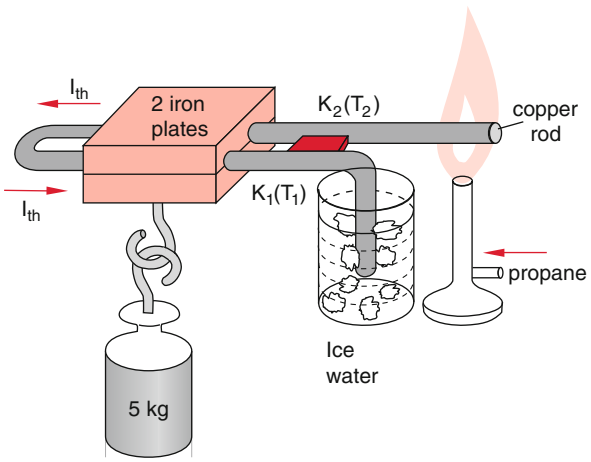


Fig. 2.68 Demonstration of the thermo-current

The relation between the Seebeck-coefficient S and thermo-current density

$$j = \sigma_{\mathrm{el}}(\nabla U - S \cdot \nabla T)$$

is

$$S = -\frac{1}{3} \frac{\mathrm{e} \cdot \Lambda \cdot n}{\sigma_{\mathrm{el}}} \frac{\mathrm{d}\bar{v}}{\mathrm{d}T} = \frac{\mathrm{e} \cdot \Lambda \cdot j_{\mathrm{ThD}}}{\sigma_{\mathrm{el}} \cdot \nabla T} \qquad (2.42h)$$

where n is the electron density, $\bar{v}$ the mean velocity of the electrons, Λ their mean free path and σ_{el} the electric conductivity (see Problem 2.15).

The thermo-voltage can be used for accurate temperature measurements (Fig. 2.67 and Vol. 1, Sect. 10.1.1) but also as voltage source for thermo-currents. This can be demonstrated with the experiment illustrated in Fig. 2.68 One end of a thick copper hoop is immersed into cold water while the other end is heated with a burner. A bar made of another material is welded between the cold and hot end of the hoop. A thermo-voltage U_{th} appears between the two ends K_1 and K_2 which causes a very large thermo-current $I_{\mathrm{th}} = U_{\mathrm{th}}/R$ (more than 100 A) through the copper hoop because of its extremely low resistance R. The current can be demonstrated by the magnetic field which it produces in two iron plates which are placed on top of each other without fixed connection. The magnetic field is so strong that the lower iron plate can carry a weight of more than 5 kg. It falls down as soon as the burner is removed.

2.9.4 Peltier-Effect

When a current is sent through a conducting bar that consists of different metals in the sequence ABA (Fig. 2.69), one of the contacts cools down while the other warms up. If the

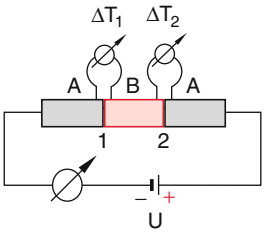


Fig. 2.69 Peltier-Effect

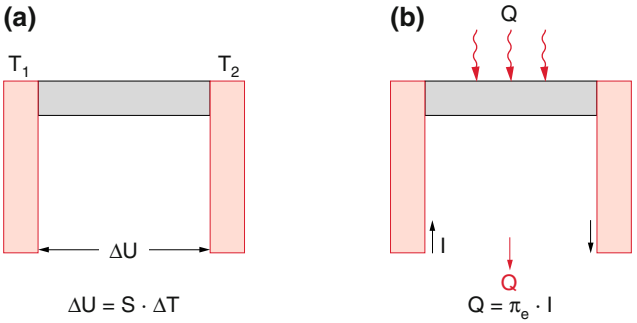


Fig. 2.70 **a**) Generation of a thermo-voltage **b**) heat transprot using the Peltier-effect

polarity of the voltage source is reversed the sign of the temperature changes ΔT_1 and ΔT_2 also have reverse polarity.

This *Peltier-effect* represents the reverse of the generation of a thermo-current. The temperature increase occurs at that contact, which is the colder one for the same direction of the thermo-current.

The heat power, produced at the contact 1 is proportional to the current I:

$$\mathrm{d}W/\mathrm{d}t = (\Pi_\mathrm{A} - \Pi_\mathrm{B}) \cdot I \tag{2.43}$$

where Π_A and Π_B are the Peltier coefficients of the materials A and B. The sign of $\mathrm{d}W/\mathrm{d}t$ depends on the direction of the current. For $\mathrm{d}W/\mathrm{d}t > 0$ heat is produced (the contact warms up), while for $\mathrm{d}W/\mathrm{d}t < 0$ heat power is extracted from the contact (it cools down). Typical numerical values of the Peltier coefficients are $\Pi \approx 10^2$ J/K. The empirical relation between thermo-voltage U_th and Peltier coefficient Π_P can be written as

$$U_\mathrm{th} = \frac{\Pi_\mathrm{e}}{T} \cdot \Delta T. \tag{2.44}$$

2.9.5 Thermo-electric Converters

Thermo-elements and Peltier-elements belong to the more general group of thermo-electric converters. These devices either produce an electric current by a temperature difference or they generate a temperature difference by sending a current through contacts between different conductors. They represent a modern research area because they can often solve problems of efficient heat transfer or they can optimize the energy balance for the thermal current production. Viewed from the energetic side thermo-electric converters transport heat, supplying electric energy or they convert heat energy into electric energy. It is, for example, possible to use industrial waste heat for the production of electric energy. Another application is the cooling of microchips or other electronic devises by very effectively dissipating the heat produced by the chips.

Such converters are applied in many technical areas. For example the heat in the exhaust of cars can be used for the production of electric energy in order to save the battery energy. A further spectacular example is the conversion of heat produced by the decay of radioactive atoms in space probes into electric energy for the supply of the electronic devices on board.

In Fig. 2.70a and b the two processes

(a) generation of a thermo-voltage
(b) transport of heat energy

are schematically illustrated where the red and the grey bars are two different materials. For metals the Seebeck coefficient is of the order of μV/K, whereas for semiconductors it is about mV/K, i.e. two to three orders of magnitude larger (Table 2.7). An example for the heat transport induced by an electric current is shown in Fig. 2.71 where the electric current is sent through an n-semiconductor and a p-semiconductor, which are connected by a metal plate. For the correct polarity of the voltage source the upper metal plate is cooled while the lower plates are heated up. This implies that heat is transported from the upper side of the semiconductors to the lower side. The energy conversion efficiency is defined by the ratio

$$\eta = \Delta Q/W_\mathrm{el} \leq (T_2 - T_1)/T_\mathrm{w} \quad \text{with} \quad T_2 > T_1. \tag{2.45}$$

of transported heat energy to supplied electric energy. Because the Carno-efficiency (see Vol. 1, Sect. 10.3) sets an upper limit to the efficiency of any device which cannot be larger than the ratio of realized temperature difference to the higher temperature.

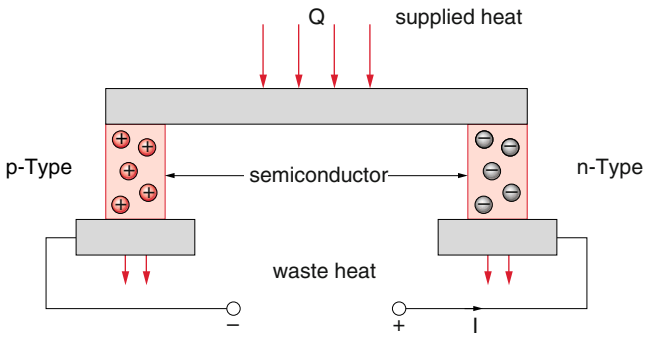


Fig. 2.71 Effective heat transport using the Peltier-effect with two different semiconductors

The maximum efficiency can be related to the efficiency number ZT, defined as

$$\mathrm{ZT} = S^2 \sigma T / \lambda \tag{2.46}$$

where S is the Seebeck coefficient, σ the electric conductivity and κ is the thermal conductivity. In Fig. 2.72 the quantities S, σ λ and ZT are plotted as a function of the free carrier charge density. The figure illustrates that semiconductors have by far the highest values of ZT. In Fig. 2.73 the temperature dependence of ZT is plotted for some semiconductors.

The efficiency η depends not only on ZT but also on the temperatures T_w of the warm side and T_c on the cold side (Fig. 2.73). The maximum achievable efficiency is (Fig. 2.74)

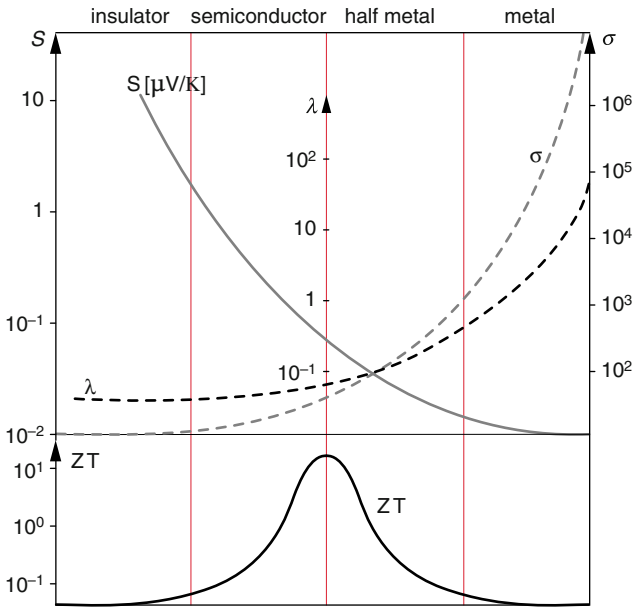


Fig. 2.72 Relations between the Seebeck coefficient S, the electric conductivity σ, the thermal conductivity λ and the energy efficiency ZT for different materials

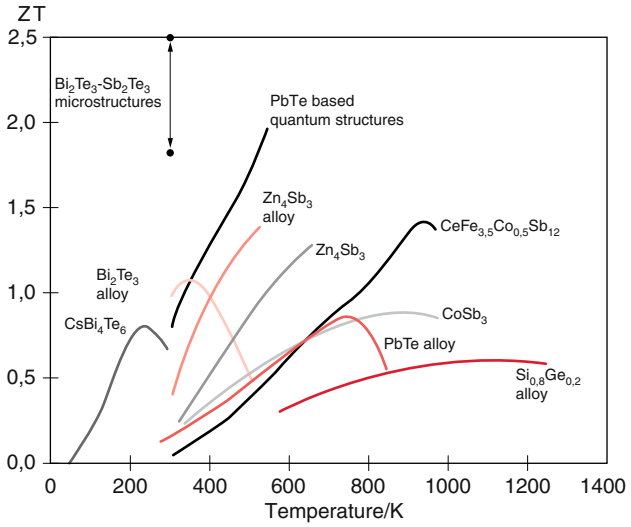


Fig. 2.73 Efficiency number $Z \cdot T$ for some semiconductor compounds as a function of temperature

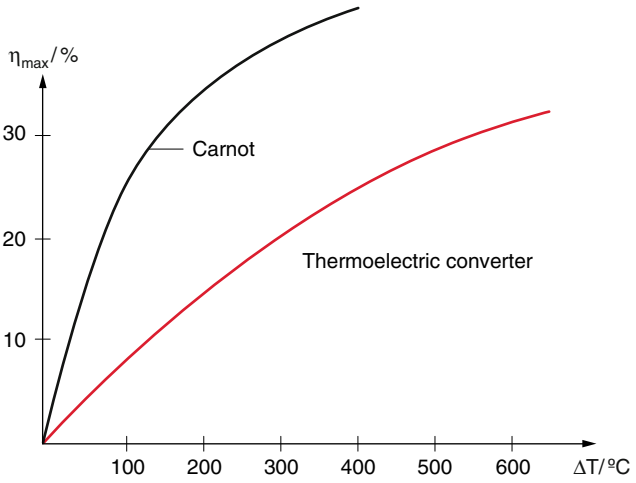


Fig. 2.74 Energy efficiency of thermoelectric converters compared to that of the Carnot-cycle

$$\eta_{\max} = \frac{T_w - T_k(\sqrt{\mathrm{ZT}+1}) - 1}{T_w(\sqrt{\mathrm{ZT}+1} - 1)}. \tag{2.47}$$

Example

$$\mathrm{ZT} = 0.6, T_w = 400\,\mathrm{K}, T_c = 300\,\mathrm{K} \rightarrow \eta_{\max} = 0.35.$$

The example shows that the maximum efficiency is not very high [18].

It is limited by the unwanted heat conduction through the Peltier element.

2.9.6 Thomson Effect

The Thomson effect (do not mistake this with the Joule-Thomson effect in Thermodynamics Vol. 1, Sect. 10.4.2) describes the altered heat conduction through a current carrying conductor if a temperature gradient exists along the conductor. The current I with the current density j produces in a conductor with specific electric resistance ρ_{el}, cross section A and length L the heat energy per second in the volume $V = A \cdot L$

$$(\mathrm{d}Q_1/\mathrm{d}t) = \varrho_{el} \cdot j^2 \cdot V \tag{2.48}$$

This heat energy is dissipated through heat conduction and is transferred to the surrounding. We regard a straight wire in the x-direction with cross section A and heat conduction coefficient λ. If we assume that there is no temperature gradient in the radial direction, we need only to consider the heat transfer into the length direction. The heat losses are then

$$\mathrm{d}Q_2/\mathrm{d}t = -\lambda \cdot A \cdot \mathrm{d}T/\mathrm{d}x \tag{2.49}$$

if the current has been switched off at $t = 0$. If the current flow continues an additional heat transport occurs with a sign that depends on the direction of the current. It can be quantitatively described by

$$\mathrm{d}Q_3/\mathrm{d}t = -\mu \cdot j \cdot \mathrm{d}T/\mathrm{d}x. \tag{2.50}$$

The energy balance in the volume element $\mathrm{d}V$ is then

$$\mathrm{d}Q/\mathrm{d}t = \varrho_{\mathrm{el}} \cdot j^2 - (\lambda \cdot A + \mu \cdot j)\mathrm{d}T/\mathrm{d}x. \tag{2.51}$$

The three phenomena *Peltier-effect*, *Seebeck effect* and *Thomson effec*t are not independent of each other. Between the corresponding coefficients the following relations exist:

$$\Pi = S \cdot T; \quad \mu = T \cdot \mathrm{d}S/\mathrm{d}T. \tag{2.52}$$

Summary

- The electric current is a transport of electric charges. It is always connected with a mass transport. The current density

$$\boldsymbol{j} = n^{+} q^{+} \boldsymbol{v}_{\mathrm{D}}^{+} + n^{-} q^{-} \boldsymbol{v}_{\mathrm{D}}^{-}$$

depends on the densities $n^{\pm}$ of the charge carriers with the charge $q^{\pm}$ and on their drift velocity $v_D^{\pm}$.

- The relation between current density and electric field strength is given by Ohm's law

$$\boldsymbol{j} = \sigma_{\mathrm{el}} \cdot \boldsymbol{E}$$

The electric conductivity σ_{el} is dependent on the material and generally also on the temperature.

- The specific electric resistance $\varrho_{\mathrm{s}} = 1/\sigma_{\mathrm{el}}$ of a conductor is caused by collisions of the charge carriers with the atoms of the conducting material. The total resistance of a conductor depends also on its geometry.
- The calculation of even complex networks is facilitated by Kirchhoff's rules, which state:

(a) At the junction of several electric conductors the sum of all currents is zero

$$\sum_k I_k = 0.$$

(b) In a closed circuit of a network of resistors or capacitances the total voltage is zero

$$\sum_k U_k = 0.$$

- In gas discharges electrons and ions both contribute to the discharge current. For the non-self-maintained discharge the discharge ends, if no longer charge carriers are generated. In stable self-maintained discharges every charge carrier has to supply its on substitute.
- In current sources energy is required to separate positive and negative charges. This spatial charge separation generates a potential difference resulting in the voltage U_0 between the poles of the current source. The source can be used as energy storage. Connecting the poles by a conductor with resistance R an electric current $I = U/(R + R_{\mathrm{i}})$ flows. The inner resistance R_{i} of the source depends on the source material and on the path length between the location of charge separation and the poles of the source.
- With an external load R_{a} the voltage drops to $U = U_0 - I \cdot R_{\mathrm{i}}$. With $I = U/R_{\mathrm{e}}$ this gives $U = U_0/(1 + R_{\mathrm{i}}/R_{\mathrm{a}})$.
- The voltage of chemical current sources is determined by the difference between the contact potentials of the two electrodes.
- The different temperature dependence of the contact potential between to different metals is used for thermometers and electric cooling (Peltier effect).

Problems

2.1 A light bulb is connected to a dc-voltage source by two 10 m long copper cables and a switch. When the switch is closed a current of 1 A flows through the cables. The density of copper is $\varrho = 8.92\,\mathrm{g/cm^3}$ and the concentration of charge carriers is $n = 5 \times 10^{28}\,\mathrm{m^{-3}}$.

(a) What is the fraction of charge carriers to the number of neutral copper atoms?

(b) If the switch is closed at $t = 0$ at which time t_1 starts the light bulb to shine? What is the time dependence $I(t)$ of the current I?

(c) Calculate the time t_2 when the first electron of the voltage source reaches the incandescent filament of the bulb? Why is t_1 so much shorter than t_2?

(d) How long must the current of 1 A flow, until 1 g electrons passes through a cross section of the filament?

2.2 A 1 m long iron wire has at one end the diameter of 1 mm and tapers uniformly to 0.25 mm at the other end. Calculate

(a) the total resistance of the wire (ρ_{el}(iron) = 8.71 $\times$ 10^{-8} Ω m)

(b) The supplied electric power per unit length, when a voltage of $U = 1$ V is applied between the ends of the wire.

2.3 Calculate the resistance between A and B in Fig. 2.75.

2.4 What are the currents I_1, I_2 and I_3 in the network of Fig. 2.76? How large is the potential difference between A and ground?

Numerical example: $U_1 = 10$ V; $R_i(U_1) = 1$ Ω; $U_2 = 4$ V; $R_{i9}(U_2) = 1$ Ω, $R_1 = 1$ Ω; $R_2 = 4$ Ω; $R_3 = 4$, $R_4 = 8$ Ω; $R_5 = 12$ Ω; $R_6 = 24$ Ω.

2.5 A car accumulator has a voltage of $U_0 = 12$ V without load. Starting the motor the current is $I = 150$ A and the voltage drops to $U_1 = 10$ V.

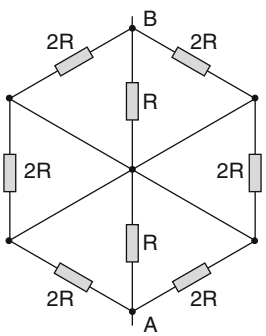


Fig. 2.75 Illustration of Problem 2.3

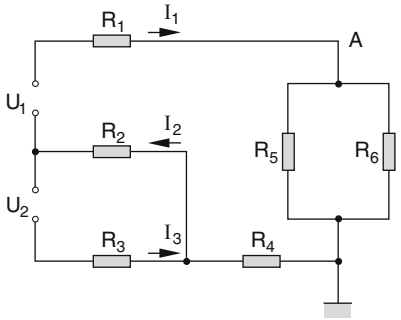


Fig. 2.76 Illustration of Problem 2.4

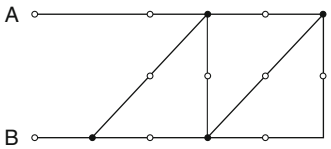


Fig. 2.77 Illustration of Problem 2.6

(a) What is the internal resistance R_i of the accumulator and of R_a = resistance of the starter?

(b) At low temperatures the internal resistance R_i increases to $R_i = R_a$. How large is now U_1?

(c) How large is under the conditions of (a) and (b) the electric power consumed in the starter and in the accumulator?

2.6 The points A and B in Fig. 2.77 are the endpoints of a circuit consisting of 8 elements (indicated by circles).

(a) What is the total capacity if the points represent equal capacitors C?

(b) What is the total resistance R if the points represent equal resistors R_i?

2.7 A cylinder of 12 cm diameter and 60 cm length is put into a nickel salt solution in order to cover it with a 0.1 thick nickel layer. The current density should not exceed 25 $\mathrm{A/m^2}$,

(a) Which maximum current is possible?

(b) What is the electro-chemical equivalent E_C? (Nickel-ions have the mass density $\rho_m = 8.7$ $\mathrm{g/cm^3}$ and the charge $q = 2e$, $m_{Ni} = 58.71$ 1.67×10^{-27} kg, the Avogadro constant is 6.023×10^{23}/mol, $e = 1.6\times10^{-19}$ C.

(c) How long have the cylinders to stay in the bath, if always the maximum current I_m flows.

2.8 A voltage source with the electro-motive force $EMF = 4.5$ V and an internal resistance $R_i = 1.2$ Ω is connected to the external load R_a. What is the

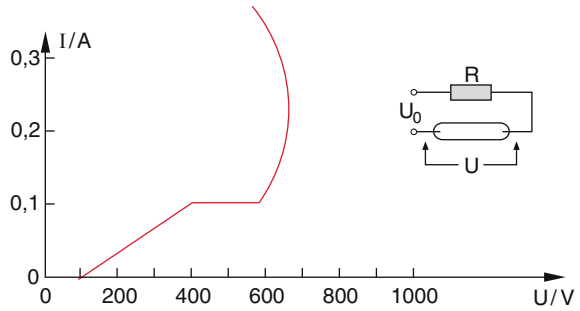


Fig. 2.78 Illustration of Problem 2.10

optimum value of R_a in order to supply the maximum power to R_a? How large is this maximum power?

2.9 A capacitor $C_1 = 20$ µF is charged to 1000 V. Then it is connected to an uncharged second capacitor $C_2 = 10$ µF, through a conducting wire with the resistance R.

(a) What were charge Q_1 and Energy W_1 of C_1 before the connection?

(b) What are voltage, total charge and total energy of $C_1 + C_2$ after the connection? Where did the energy difference go?

2.10 Assume the current-voltage characteristics of a gas discharge as that shown in Fig. 2.78.

(a) Calculate the values R_{max} and R_{min} for the dropping resistor to achieve a stable discharge when a voltage of 1000 V is applied.

(b) Assume the dropping resistor is $R = 5$ kΩ. Which changes occur in the discharge if the voltage is changed to 500 V resp. 1250 V?

2.11 Assume a KCl solution has the electric conductivity $\sigma_{el} = 1.1\ (\Omega\ m)^{-1}$. What are the amplitudes of the alternating ion motion in an electric ac-field with $E = 30$ V/cm and a frequency of $f = 50\ s^{-1}$ for an ion density $n^+ = n^- = 10^{20}/cm^3$ and equal mobilities $u^+ = u^-$?

2.12 A shielded cable consisting of an inner conductor ($r_1 = 1$ mm) and concentric metallic bush (inner radius $r_2 = 8$ mm) is filled with isolating material ($\varrho_s = 10^{12}\,\Omega\,m$). How large is the leakage current through the isolating material between inner and outer conductor for a cable length of 100 m and a voltage of 3 kV between inner and outer conductor?

2.13 The cable described in 2.12 can be represented by the circuit shown in Fig. 2.79, where R_1 is the resistance of the cable per m and R_2 is the leakage resistance per m.

(a) What is the resistance R_n between a and b for n meter cable length?

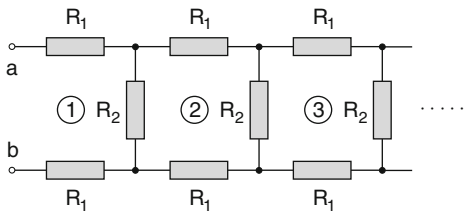


Fig. 2.79 Illustration of Problem 2.13

(b) How large is for $R_1 = R_2$ the limit $\lim R_n$ for $n \to \infty$?

2.14 Derive the relation (2.42d).

2.15 Why illustrates Eq. (2.42h) that the thermo-voltage is not caused by the contact potentials? Derive (2.42h) (Note: you can find help in [17]).

References

1. CRC handbook of Chemistry and Physics Vol. 99 (2018).
2. https://en.wikipedia.org/wiki/Heike_Kamerlingh_Onnes.
3. https://en.wikipedia.org/wiki/High-temperature_superconductivity.
4. E.W. Carlson, V.J. Emery, S.A. Kivelson, D. Orgard: Concepts in High Temperature Superconductivity. In: The physics of Conventional and Unconventional Superconductors ed., by K.H. Benneman et al. (Springer, https://www.physics.purdue.edu/~erica/talks/Concepts.pdf).
5. V.V. Schmidt, The physics of superconductors: Introduction to fundamentals and applications. Springer Science & Business Media, 2013.
6. V.L. Ginzburg: On Superconductivity and Superfluidity (Springer 2008).
7. J. Swineburne: The Measurements of electric Currents (Palala Press 2016).
8. https://en.wikipedia.org/wiki/Electric_current.
9. H. Stöcker: Taschenbuch der Physik (Verlag Harri Deutsch Frankfuret 1994).
10. B.M. Smirnov; Theory of Gas Discharge Plasma (Springer Series on Atomic, Optical and Plasma Physics Vol. 84. (Springer Heidelberg 2015); Y.P.Raizer: Gas Discharge Physics (Springer Heidelberg 1991).
11. Linden, David; Reddy, Thomas B., eds. (2002). Handbook Of Batteries (3rd ed.). New York: McGraw-Hill. p. 23.5. ISBN: 978-0-07-135978-8.
12. Electric batteries books group ed. (books LLC 2010).
13. Beta Writer: Lithium Ion Batteries (Springer 2019); C.Julien, A. Mauger, A.Vijh, K.Zaghib: Lithium Batteries (Springer 2016).
14. https://www.explainthatstuff.com/fuelcells.html.
15. https://en.wikipedia.org/wiki/Fuel_cell.
16. M. Cultu:Batteries and their Chemistry. Energy storage systems in: Encyclopedia of life support systems,(UNESCO EOLSS).
17. https://en.wikipedia.org/wiki/Thermoelectric_effect.
18. https://en.wikipedia.org/wiki/Thermoelectric_generator.

Static Magnetic Fields

3

Already in ancient times it was observed that special minerals that were found around the city *Magnesia* in Asian Turkey, attracted iron matter. They were named *magnets* and were used for producing compass needles for navigation, because it was found that such needles always pointed to the north. The Chinese knew about magnets already earlier. The exact explanation of such magnetic properties had, however, to wait until the 20th century after the development of quantum theory and modern solid state physics. There are still open questions about the details of magnetic phenomena in matter.

In Sect. 2.5 we have discussed that also electric currents show magnetic effects.

In the present chapter we discuss in more detail the magnetic fields produced by permanent magnets and by electric currents. The properties of magnetic materials will be here treated only phenomenological, while in Vol. 3 of this series it will be shown that also the magnetic properties of matter are caused by atomic electric moments and electric currents on an atomic scale.

3.1 Permanent Magnets

We will start with some basic experiments.

When iron powder is spread onto a glass plate, where a permanent bar magnet is placed below the glass plate, one observes that the iron powder arranges in form of lines which accumulate above two points of the bar magnet (Fig. 3.1). We call these accumulation points the ***magnetic poles***.

If a rod shaped permanent magnet is suspended by a wire at its center of mass, it can turn around its center of mass. One observes that one pole is alway pointing to the north (we call this pole therefore the *magnetic north pole*, the other pole points towards south (*magnetic south pole*). When a second magnetic rod is neared to the turnable rod (Fig. 3.2) the north pole of one rod is attracted by the south pole of the other rod, but repelled by the north pole.

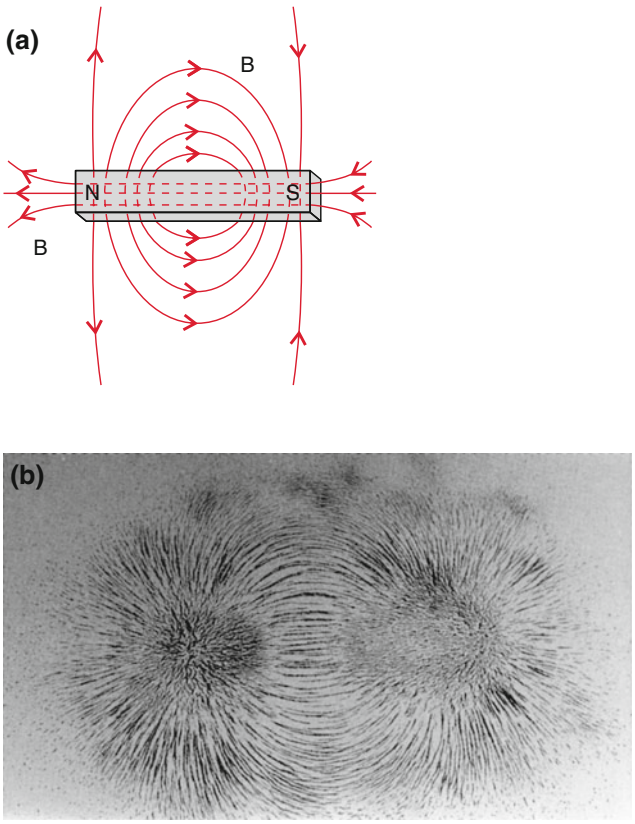


Fig. 3.1 **a**) Magnetic field lines of a magnetic rod, **b**) experimental demonstration of magnetic field lines with iron powder. *Note* that the field lines form closed curves i.e. they continue inside the magnetic rod

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_3

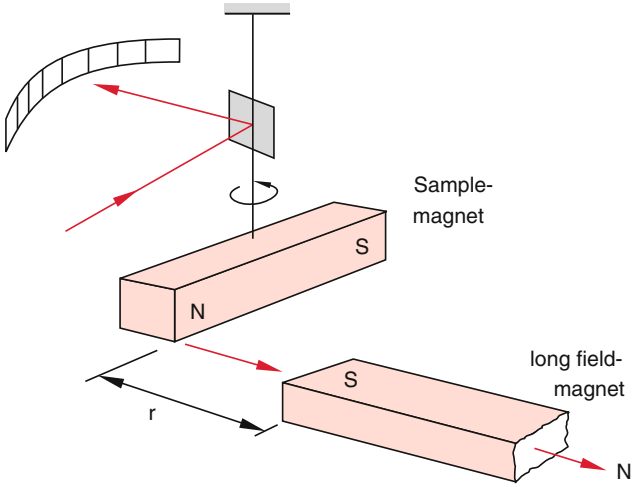


Fig. 3.2 Magnetic torsion balance for measuring the force $F(r)$ between magnetic poles. *Note* The distance between the poles of the field magnet must be large compared with the distance between the poles to be measured

Equal poles repel each other, while opposite poles attract each other. This is completely analogous to the situation in electrostatics where equal charges repel each other while opposite charges attract each other.

However, there is a fundamental difference: If a magnetic rod is broken in the mid into two halfes there will be no separated poles but each piece is again a magnet with north and south pole (Fig. 3.3). One can continue this process of breaking each piece again into two halfes and both parts represent again a magnet with north- and south pole. From this experiment we can conclude:

There are no isolated magnetic poles. Always only dipoles exist, no monopoles.

Another difference between electric and magnetic fields shall be emphasized: Electric field lines start at the positive charge and end at the negative charge, whereas magnetic field lines are always closed lines. The impression, that they apparently start at the north pole and end at the south pole is not correct, because they continue inside the magnetic rod from the south- to the north pole and thus form closed lines (compare Figs. 1.10 and 3.1).

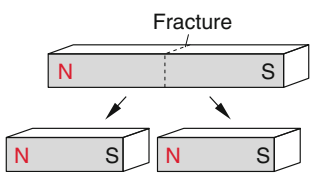


Fig. 3.3 Dividing a magnetic rod into two pieces does not separate individual poles but creates two dipole magnets

It will be shown later that static *electric* fields are generated by charges at rest, whereas static *magnetic* fields are produced by moving charges, i.e. by electric currents.

Figure 3.1 shows that magnetic fields can be illustrated by field lines analogous to electric fields. They give for every point in space the direction of the field as tangent to the field lines. We can, as for electric fields, define a magnetic flux

$$\Phi_m = \int \boldsymbol{B} \cdot \mathrm{d}A \tag{3.1}$$

through the area A, which can be illustrated by the number of field lines through A. The quantity $\boldsymbol{B}$ [V s/m^2] is the *magnetic field strength* (often called the magnetic flux density). For abbreviation purposes a new unit for the magnetic field B is introduced:

$$1\,\text{Tesla} = 1\,\text{T} = \;= 1\,\text{V s/m}^{-2} \tag{3.2}$$

For practical purposes 1 T is a very large unit. Therefore smaller units are in use:

$$1\,\text{mT} = 10^{-3}\,\text{T}$$

$$1\,\mu\text{T} = 10^{-6}\,\text{T}$$

$$1\,\text{Gauss} = 1\,\text{G} = 10^{-4}\,\text{T}$$

Compare: The unit of the electric flux Φ_{el} is V m, that of the electric field strength E is V/m, the magnetic flux Φ_m has the unit V s, whereas the unit of the magnetic field strength $\boldsymbol{B}$ is V s/m^2.

Examples

The average magnitude of the earth magnetic field is 20 μT = 0.2 G. With large superconductive magnets one can reach magnetic fields up to 30 T. With hybrid magnets, where in addition to the magnetic field of the superconductive coils a second field is superimposed, produced by coils with normal conduction, a field strength up to 40 T can be realized.

3.2 Magnetic Fields of Stationary Currents

When an electric current I is sent through a long straight wire one observes that a compass needle in the distance r from the wire is deflected in such a way, that it always points in the direction of the tangent to a circle with radius r around the wire (Fig. 3.4). This demonstrates that the electric current produces a magnetic field. It can be visualized by iron powder spread on a sheet around the wire. Placing small

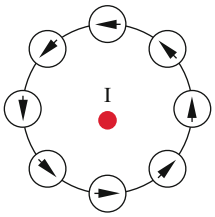


Fig. 3.4 Measuring the magnetic field of a current carrying wire with a small magnetic needle

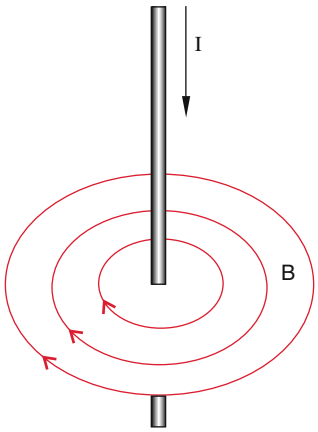


Fig. 3.5 Magnetic field lines around a straight current carrying wire

compass needles on a circle around the wire (Fig. 3.4) shows the direction of the magnetic field. Viewing into the direction of the electric current I the magnetic field B on the circle is directed clockwise (Fig. 3.5). A convenient rule is: If the thumb of the right hand points into the direction of the current I the curved other fingers give the magnetic field lines and the direction of B.

A current through a solenoid (Fig. 3.6) produces a magnetic field which is similar to that of a magnetic rod. If such a solenoid is fixed to a torsion balance as in Fig. 3.2, a completely equivalent behavior as for a bar magnet is observed: One end of the solenoid acts like a north pole, the other as a south pole. Reverting the current through the solenoid also interchanges the magnetic poles. In a solenoid the magnetic field lines can be visualized also inside the solenoid (Fig. 3.6). This illustrates that the magnetic field lines are closed lines which do not end at the poles.

In this chapter we will show, how magnetic fields of arbitrary arrangements can be calculated. For this goal we have to introduce some new terms and definitions.

3.2.1 Magnetic Flux and Magnetic Voltage

Since the magnetic lines are closed lines we can conclude that the magnetic flux through a closed surface (Fig. 3.7) must be always zero, because there are as many lines entering the

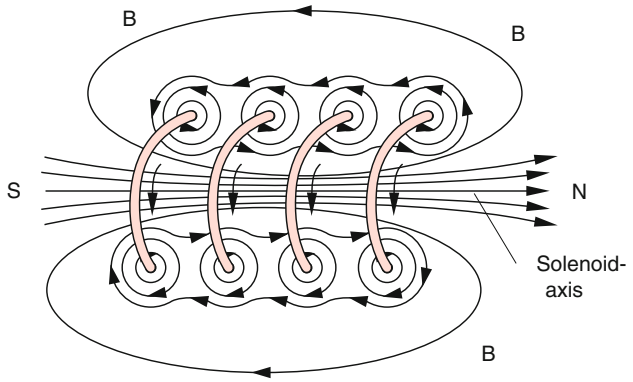


Fig. 3.6 Magnetic field of a long current carrying solenoid

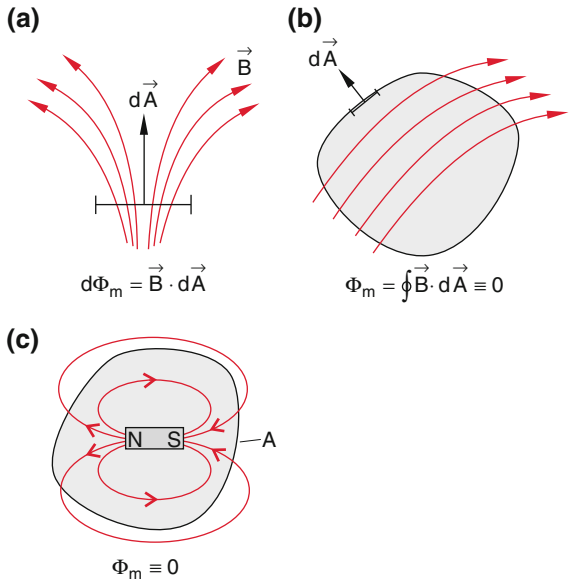


Fig. 3.7 The magnetic flux through a closed surface is zero

volume enclosed by the surface, as lines traversing the surface from the inside. In a mathematical form this means

$$\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A} = 0 \tag{3.3}$$

Transformation of this surface integral into a volume integral over the enclosed volume yields according to Gauss' law:

$$\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A} = \int \mathrm{div}\boldsymbol{B}\, \mathrm{d}V \equiv 0$$

Since this is valid for arbitrary closed surfaces it follows that

$$\mathrm{div}\, \boldsymbol{B} = 0 \tag{3.4}$$

This is the mathematical formulation for the fact that no magnetic monopoles exist. Sources and sinks of the

magnetic field (north- and south poles) always exist in pairs, contrary to the electric field where isolated charges of one sign can exist and where for a charge density ρ the divergence of the electric field is

$$\operatorname{div} \boldsymbol{E} = \varrho/\varepsilon_0 \neq 0.$$

In the electrostatic field the line integral

$$\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} = U$$

is equal to the electric voltage $U = \phi_1 - \phi_2$ between two points with the potentials ϕ_1 and ϕ_2. Integrated over a closed path is $\int \boldsymbol{E} \cdot \mathrm{d}\boldsymbol{s} = 0$. On the other side for the magnetic field the integration over a closed path is **not** zero!

The experiments give for a closed path around a wire or an area with current I

$$\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{s} = \mu_0 \cdot I, \qquad \textbf{Ampere's law} \tag{3.5}$$

The constant

$$\mu_0 = 4\pi \times 10^{-7} \frac{\mathrm{V\,s}}{\mathrm{A\,m}} \tag{3.6}$$

is called **magnetic induction constant**. From (3.5) we can then obtain the unit of $\boldsymbol{B}$ as [V s/m^2].

Because the current density j of a current I through the area A is related to I by

$$I = \int \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A}$$

we can transform (3.5), using *Stokes Theorem* into

$$\mu_0 \int \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A} = \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{s} = \int \mathbf{rot}\,\boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A}.$$

Since this is valid for arbitrary integration paths it follows:

$$\mathbf{rot}\,\boldsymbol{B} = \mu_0 \cdot \boldsymbol{j} \tag{3.7}$$

whereas for the electric field is $\mathbf{rot}\,\boldsymbol{E} = 0$ (see (1.65c)).

Static electric fields are curl-free, in contrary to static magnetic fields.

Using Ampere's law and the magnetic flux Φ_m we can readily calculate the magnetic fields of some special current distributions. This will be demonstrated in the next sections.

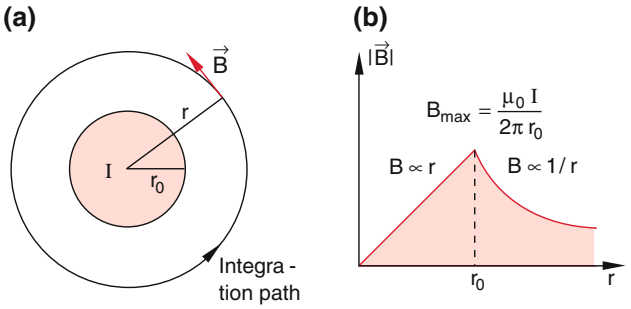


Fig. 3.8 **a**) Integration path along the circular magnetic field line around a straight current carrying wire, **b**) magnetic field strength $B(r)$ as the function of the distance r from the wire center $r = 0$

3.2.2 The Magnetic Field of a Straight Cylindrical Conductor

Experiments measuring the magnetic field of a straight wire carrying the current I (Figs. 3.5 and 3.6) have proved that the magnetic field lines are concentric circles with radius r where on each circle $B(r)$ = const. We choose as integration path such a circle with radius r around the conductor with radius r_0 (Fig. 3.8a). For $r > r_0$ this gives in polar coordinates

$$\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{s} = \int r \cdot B \cdot \mathrm{d}\varphi = 2\pi r \cdot B(r) = \mu_0 \cdot I.$$

The amount of B is then

$$B(r) = \frac{\mu_0 I}{2\pi r}. \tag{3.8}$$

For $r < r_0$ only the fraction $\pi r^2 \cdot j$ of the total current I is enclosed by the integration path. We now obtain:

$$\begin{aligned} 2\pi r \cdot B(r) &= \mu_0 \pi r^2 \cdot j \\ \Rightarrow B(r) &= \frac{1}{2}\mu_0 \cdot j \cdot r = \frac{\mu_0 \cdot I}{2\pi r_0^2} r. \end{aligned} \tag{3.9}$$

$B(r)$ has its maximum value at the surface ($r = r_0$) of the conductor (Fig. 3.8b).

3.2.3 Magnetic Field in the Inside of a Long Solenoid

The experimental illustration with iron powder shows that the magnetic field is concentrated in the inside of the solenoid where it is practically homogeneous (Fig. 3.6) whereas outside the solenoid it is very weak and nearly negligible, if the

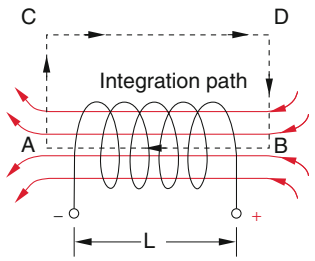


Fig. 3.9 Integration path for the determination of the magnetic field of a long solenoid

diameter of the solenoid is small compared to its length L. We choose the integration path as shown by the dashed line in Fig. 3.9. Since only the path inside the solenoid gives a noticeable contribution (on the way AC and DB is $\boldsymbol{B} \perp \mathrm{d}\boldsymbol{s}$ and outside the solenoid we can place the path far away from the solenoid, where the magnetic field B is very small) we get

$$\int \mathbf{B} \cdot \mathrm{d}s \approx \int_B^A B \, \mathrm{d}s = B \cdot L = N \cdot \mu_0 \cdot I \\ \Rightarrow B = \mu_0 n \cdot I \tag{3.10}$$

where $n = N/L$ is the number of windings per m and L is the length of the solenoid. This simplified treatment gives the result, that the magnetic field is homogeneous in the inside, i.e. independent of the location.

Example

$N = 10^3/\mathrm{m}$, $I = 10\,\mathrm{A}$, $\mu_0 = 1.26 \cdot 10^{-6}\,\mathrm{V\,s(A\,m)}$
$\Rightarrow B = 0.0126\,\mathrm{T} = 126\,\mathrm{G}$

3.2.4 Vector Potential

In the Sects. 1.3 and 1.4 it was shown, that there is a general method to calculate the electrostatic potential $\phi(\boldsymbol{r})$ using Eq. (1.20) and the electric field $\boldsymbol{E}(\boldsymbol{r}) = -\mathbf{grad}\,\phi(\boldsymbol{r})$ either analytically or at least numerically, if the charge distribution $\rho(r)$ is known. The question now arises whether the same is true for magnetic fields, i.e. whether the magnetic field $\boldsymbol{B}(\boldsymbol{r})$ and a "magnetic potential", which still has to be defined, can be calculated, if the distribution of electric currents is known.

We learn from Eq. (3.6) that $\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{s} \neq 0$ when the integration path encloses electric currents. In such cases the integral $\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{s}$ is no longer independent of the integration path. Therefore it is not possible to define a magnetic potential ϕ_m that obeys the relation $\boldsymbol{B} = -\mu_0 \cdot \mathbf{grad}\,\phi_m$ as in the electrostatic case (see Sect. 1.3), because then we would get $\mathbf{rot}\,\boldsymbol{B} = -\mu_0 \cdot \nabla \times \nabla\phi_\mathrm{m} = 0$ contrary to Eq. (3.7).

Note For the vector relation $\nabla \times \nabla = 0$ (see citation of Vol. 1, Chap. 13, or any book about vector analysis [1]).

Since div $\boldsymbol{B} = 0$ one can define a vector field $\boldsymbol{A}(\boldsymbol{r})$ by the relation

$$\boldsymbol{B} = \mathbf{rot}\,\boldsymbol{A} \tag{3.11}$$

which is called the vector potential of the magnetic field $\boldsymbol{B}$. This definition automatically fulfills the condition div $B = 0$, because

$$\mathrm{div}\,\boldsymbol{B} = \nabla \cdot (\nabla \times A) \equiv 0$$

The definition (3.11) does not determine $A(\boldsymbol{r})$ unambiguously, because another vector field

$$\boldsymbol{A}' = \boldsymbol{A} + \mathbf{grad}\,f$$

With an arbitrary scalar function $f(\boldsymbol{r})$ satisfy also (3.11) because rot grad $f \equiv 0$. This means that $\boldsymbol{A}'$ gives the same magnetic field $\boldsymbol{B}$ as A. It is therefore necessary to define an additional condition (**gauge condition**) which for static fields read:

$$\mathrm{div}\,\boldsymbol{A} = 0 \quad \text{(Coulomb Gauge)} \tag{3.12}$$

These two definitions determine A unambiguously apart from a scalar function $f(r)$ with grad $f = 0$. We can choose $f(r)$ in such a way, that

$$\boldsymbol{A}(\mathrm{r} = \infty) = 0.$$

The two definitions of A are therefore:

$$\mathbf{rot}\,\boldsymbol{A} = \boldsymbol{B} \text{ and } \mathrm{div}\,\boldsymbol{A} = 0.$$

remark Unfortunately the vector potential $\boldsymbol{A}$ and the area $\boldsymbol{A}$ are labeled with the same letter A. It should, however, not cause confusion.

3.2.5 The Magnetic Field of an Arbitrary Distribution of Electric Currents; Biot-Savart Law

In this section we will show that the vector potential $\boldsymbol{A}(\boldsymbol{r})$ of an arbitrary distribution of electric currents with current density $j(r)$ can be obtained in a completely analogous way as the electric potential $\phi_\mathrm{e}(\mathrm{r})$ from the electric charge distribution $\rho_\mathrm{el}(r)$.

From (3.7) and (3.11) we obtain with

$$\mathbf{rot}\,\boldsymbol{B} = \nabla \times (\nabla \times \boldsymbol{A}) = \mathbf{grad}\,\mathrm{div}\,\boldsymbol{A} - \mathrm{div}\,\mathbf{grad}\,\boldsymbol{A} = \mu_0 \boldsymbol{j}.$$

Because $\mathrm{div}\,\boldsymbol{A} = 0$ we obtain with $\mathrm{div}\,\mathbf{grad}\,\boldsymbol{A} = \Delta\boldsymbol{A}$ (Δ is the Laplace operator) the relation

$$\Delta \boldsymbol{A} = -\mu_0 \cdot \boldsymbol{j} \tag{3.13}$$

Writing (3.13) for the three components this becomes

$$\Delta A_i = -\mu_0 \cdot j_i, \quad i = x, y, z \tag{3.13a}$$

Note that these three component equations are completely analogous to the Poisson equation $\Delta\phi_{\mathrm{el}} = -\rho/\varepsilon_0$ when one replaces the current density j by the charge density ρ and μ_0 by $1/\varepsilon_0$.

Therefore also the solutions must be equivalent. Analogous to (1.20) we get for the vector potential $\boldsymbol{A}(\boldsymbol{r}_1)$ at the point $\boldsymbol{P}(\boldsymbol{r}_1)$ the vector equation

$$\boldsymbol{A}(\boldsymbol{r}_1) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{j}(\boldsymbol{r}_2)\mathrm{d}V_2}{r_{12}} \tag{3.14}$$

with $r_{12} = |\boldsymbol{r}_1 - \boldsymbol{r}_2|$. The integration extends over the whole volume of electric currents (Fig. 3.10).

When the vector potential of a given current distribution has been calculated, the magnetic field $\boldsymbol{B}(\boldsymbol{r}_1)$ in the observation point $\boldsymbol{r}_1$ can be obtained from $\boldsymbol{B} = \mathbf{rot}\,\boldsymbol{A}$ by differentiation with respect to the coordinates $\boldsymbol{r}_1$ of the reference point $\boldsymbol{P}(\boldsymbol{r}_1)$.

Note, that the differentiation must be performed with respect to the reference point $P(r_1)$, but the integration has to be performed over the volume $\mathrm{d}V_2$ of the current area. The succession of differentiation and integration can be interchanged. This yield for the magnetic field:

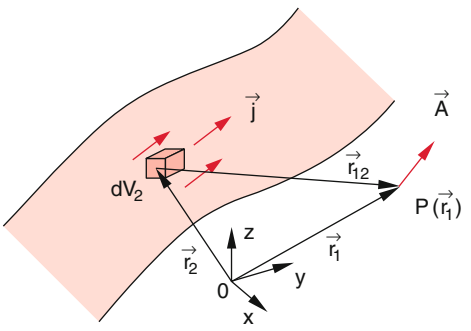


Fig. 3.10 Vector potential $\boldsymbol{A}(\boldsymbol{r}_1)$ of the current distribution $\boldsymbol{j}(\boldsymbol{r}_2)$

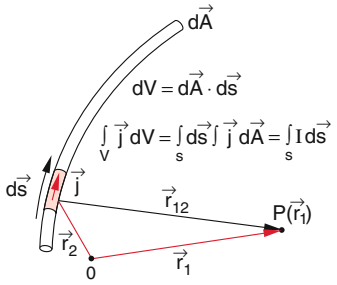


Fig. 3.11 Illustration of the Biot-Savart law

$$\boldsymbol{B}(\boldsymbol{r}_1) = \frac{\mu_0}{4\pi} \int \nabla \times \frac{\boldsymbol{j}(\boldsymbol{r}_2) \cdot \mathrm{d}V_2}{r_{12}}. \tag{3.15}$$

With $r_{12} = \sqrt{\left[(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2\right]}$ the differentiation gives (see Problem 3.8)

$$\boldsymbol{B}(\boldsymbol{r}_1) = \frac{\mu_0}{4\pi} \int \frac{j(\boldsymbol{r}_2) \times \hat{\boldsymbol{e}}_{12}}{\boldsymbol{r}_{12}^2} \mathrm{d}V_2 \tag{3.16}$$

with the unit vector $\hat{\boldsymbol{e}}_{12} = \boldsymbol{r}_{12}/r_{12}$.

When the current only flows through thin wires (Fig. 3.11) we can simplify the integral. Because now the current density is nearly constant across the cross sectional area A we get: $\boldsymbol{j} \cdot \mathrm{d}V = \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A} \cdot \mathrm{d}\boldsymbol{s} = \boldsymbol{I} \cdot \mathrm{d}\boldsymbol{s}$ and we can immediately perform the integration over $\mathrm{d}\boldsymbol{A}$. This reduces the volume integral to a line integral

$$B(r_1) = -\frac{\mu_0}{4\pi} \cdot I \cdot \int \frac{\hat{e}_{12} \times \mathrm{d}s}{r_{12}^2} \tag{3.16a}$$

Relation (3.16a) is called **Biot-Savart Law**.

We will illustrate its application by some examples.

3.2.5.1 The Magnetic Field of a Straight Conductor

We regard in Fig. 3.12 a long straight wire in which the current flows into the +z-direction. The unit vector $\hat{\boldsymbol{e}}_{12} = \hat{\boldsymbol{e}}_r$ points from the line element $\mathrm{d}z$ to the reference point P on a

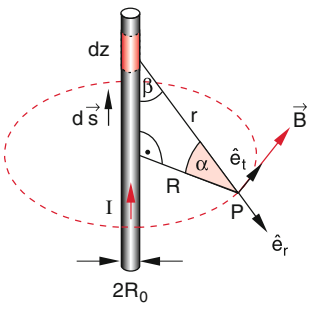


Fig. 3.12 Calculation of magnetic field and vector potential of a long straight wire

circle with radius R around the wire. The vector product $\hat{e}_{12} \times \mathrm{d}s$ has the amount

$$|\hat{e}_{12} \times \mathrm{d}s| = \sin\beta \cdot \mathrm{d}z = \cos\alpha \cdot \mathrm{d}z$$

Its direction is the tangent to the circle with radius R.

According to (3.16a) we obtain for the magnetic field $B(R)$ in the point P on the circle

$$B(R) = \frac{\mu_0 I}{4\pi}\hat{e}_t \cdot \int \frac{\cos\alpha}{r^2}\mathrm{d}z. \tag{3.16b}$$

With $r = R/\cos\alpha$, $z = R \cdot \tan\alpha \Rightarrow \mathrm{d}z = R\,\mathrm{d}\alpha/\cos^2\alpha$ we get for the amount $B = |\boldsymbol{B}|$

$$B = \frac{\mu_0 I}{4\pi R}\int_{-\pi/2}^{+\pi/2} \cos\alpha\,\mathrm{d}\alpha = \frac{\mu_0 I}{2\pi R}, \tag{3.17}$$

which we had already derived by another way in Sect. 3.2.2.

Since the current density j has only a z-component, the vector potential can also only have a z-component

$$A = \{0, 0, A_z\}.$$

With B = rot A it follows

$$B_x = \frac{\partial A_z}{\partial y}, \quad B_y = -\frac{\partial A_z}{\partial x} \quad \text{and} \quad B_z = 0.$$

In cylindrical coordinates (r, φ, z) this reads

$$B_r = \frac{1}{r}\frac{\partial A_z}{\partial \varphi} \quad \text{and} \quad B_\varphi = -\frac{\partial A_z}{\partial r}.$$

Because of the cylinder-symmetry A_z does not depend on φ. Therefore is $\partial A_z/\partial\varphi = 0 \Rightarrow B_r = 0$. For $r = R$ we then obtain

$$B = B_\varphi = -\left(\frac{\partial A_z}{\partial r}\right)_R = \frac{\mu_0 I}{2\pi R}. \tag{3.18a}$$

Integration yields for $R > R_0$

$$A_z = -\int_{R_0}^{R} B\,\mathrm{d}R = -\frac{\mu_0 \cdot I}{2\pi}\ln\frac{R}{R_0}. \tag{3.18b}$$

The boundary condition $A(\infty) = 0$ can not be applied here, similar to the potential of the charged rod in Eq. (1.18c). Since Eq. (3.18b) shows that $A_z(R_0) = 0$, we choose $R = R_0$ as the zero location for $A(R)$. In the inside of the conductor ($R < R_0$) is the current

$$\begin{aligned} I = j \cdot \pi R^2 \Rightarrow B_\varphi &= -\frac{\partial A_z}{\partial R} = -\frac{1}{2}\mu_0 j \cdot R \\ \Rightarrow A_z &= -\frac{1}{4}\mu_0 j R^2 + \text{const.} \end{aligned} \tag{3.18c}$$

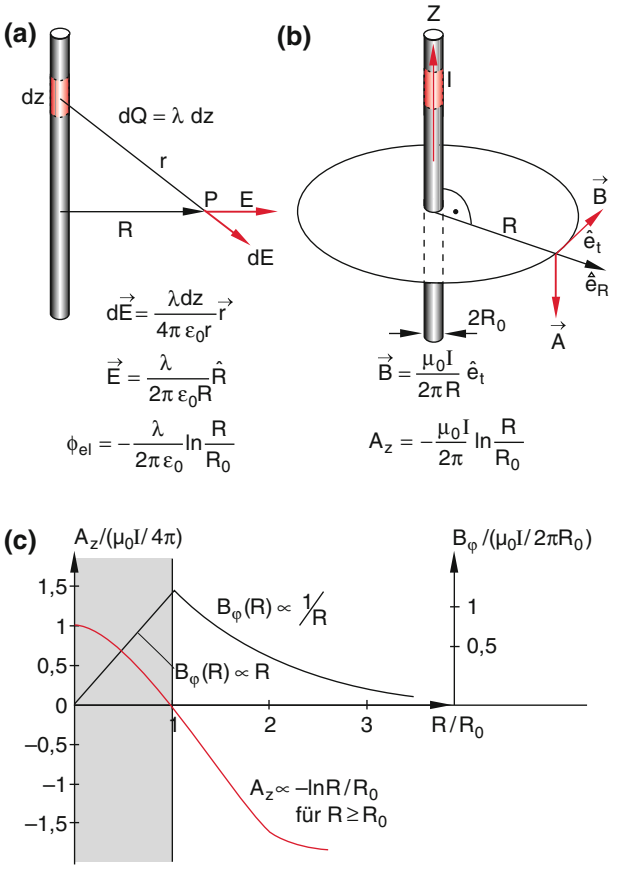


Fig. 3.13 Comparison between electric potential $\phi(R)$ of a long straight wire with radius R_0 and electric line charge density $\mathrm{d}Q/\mathrm{d}z = \lambda$ **a**) and the vector potential $A_z(R)$ of a current I carrying wire **b**). **c**) Radial dependence of $B_\varphi(R)$ and $A_z(R)$

From the condition $A_z(R_0) = 0$ the integration constant can be determined. It is

$$\text{const.} = \frac{1}{4}\mu_0 j R^2$$

The vector potential inside the conductor is then

$$A_z(R \leq R_0) = \frac{1}{4}\mu_0 j\left(R_0^2 - R^2\right). \tag{3.18d}$$

Since $I = j\pi R_0^2$ is the total current through the conductor, we can write (3.18d) as

$$A_z(R \leq R_0) = +\frac{1}{4\pi}\mu_0 I\left(1 - (R/R_0)^2\right). \tag{3.18e}$$

In Fig. 3.13c the magnetic field $|B(R)|$ and the vector potential $A_z(R)$ are plotted inside and outside the conductor. In Fig. 3.13a, b the comparison between the electric field of a charged wire and the magnetic field of the current through a straight wire is illustrated. This demonstrates the close resemblance between the two situations.

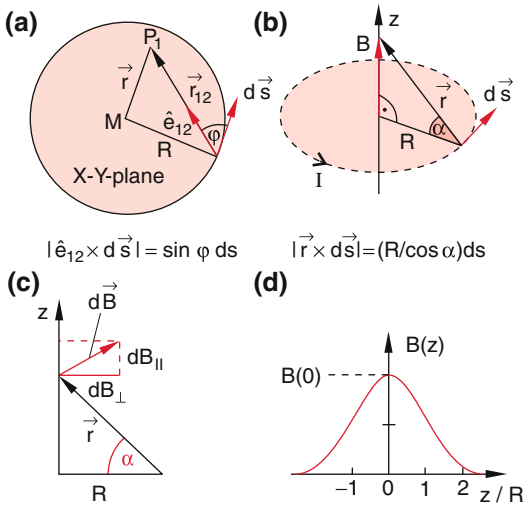


Fig. 3.14 Illustration of the calculation of the magnetic field of **a**) circular loop, **b**) on the symmetry axis, **c**) definition of $dB_\perp$ and $dB_\parallel$, **d**) Magnetic field on the symmetry axis (z-axis)

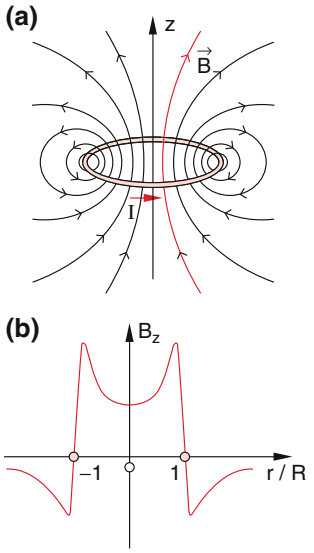


Fig. 3.15 **a**) Magnetic field lines of a circular loop, **b**) the r-dependence of $B_z(r/R)$ in the loop area

3.2.5.2 The Magnetic Field of a Circular Current Loop

When we place the current loop in the x-y-plane (Fig. 3.14a), the magnetic field B in this plane has, according to the Biot-Savart Law (3.16a) only a z-component. Its amount at the point $P_1(x, y, 0)$ is with $\hat{\boldsymbol{e}}_{12} \times d\boldsymbol{s} = \sin\varphi\, ds$

$$B_z = \frac{\mu_0 \cdot I}{4\pi} \cdot \int \frac{\sin\varphi}{r_{12}^2} ds \tag{3.19}$$

In the center of the circle is $r_{12} = R$ and $\varphi = \pi/2$. The magnetic field is then

$$B_z = \frac{\mu_0 \cdot I}{2 \cdot R}. \tag{3.19a}$$

On the symmetry axis (z-axis through the center of the circle) (Fig. 3.14b) we obtain from (3.16a) the amount $d\boldsymbol{B}(z)$ of the magnetic field $\boldsymbol{B}$ produced by the line element ds of the loop

$$d\boldsymbol{B} = -\frac{\mu_0 \cdot I}{4\pi} \cdot \frac{\boldsymbol{r} \times d\boldsymbol{s}}{r^3}. \tag{3.19b}$$

When integrating over all line elements of the circle the components $dB_\perp = dB \cdot \sin\alpha$ perpendicular to the symmetry axis average to zero. Only the parallel component $dB_\parallel = dB \cdot \cos\alpha$ remains. It yields after integration with $|\boldsymbol{r} \times d\boldsymbol{s}| = R/\cos\alpha$

$$B_\parallel = B_z = \int |dB_{||}| = \int |dB| \cdot \cos\alpha.$$

Inserting (3.19b) gives

$$B_z = \frac{\mu_0 \cdot I}{4\pi \cdot r^3} \cdot \int R \cdot ds = \frac{\mu_0 \cdot I \cdot R}{4\pi \cdot r^3} \cdot 2\pi \cdot R.$$

With $r^2 = R^2 + z^2$ this yields for the magnetic field on the symmetry axis of the loop (z-axis)

$$B_z(z) = \frac{\mu_0 \cdot I \cdot \pi \cdot R^2}{2\pi (z^2 + R^2)^{3/2}}. \tag{3.19c}$$

The z-dependence of the magnetic field on the z-axis is illustrated in Fig. 3.14d.

The magnetic field lines of the current loop are shown in the upper part of Fig. 3.15. The progression of the lines is similar to that of a short magnetic rod (Fig. 3.1).

The plane current loop represents a magnetic dipole. With the area vector $\boldsymbol{A} = \pi \boldsymbol{R}^2 \hat{\boldsymbol{e}}_z$ perpendicular to the plane with the amount giving the area enclosed by the loop, we can write the magnetic field (3.19c) as

$$\boldsymbol{B} = \frac{\mu_0}{2\pi} \frac{\boldsymbol{I} \cdot \boldsymbol{A}}{r^3}. \tag{3.20}$$

where r is the distance of the observation point on the z-axis from the loop plane.

The product

$$\boldsymbol{p}_\mathrm{m} = \boldsymbol{I} \cdot \boldsymbol{A} \tag{3.21}$$

of electric current $\boldsymbol{I}$ and area vector $\boldsymbol{A}$ is the **magnetic dipole moment** $\boldsymbol{p}_\mathrm{m}$ of the current loop.

Inserting (3.21) into (3.20) we can express the magnetic field on the symmetry axis of the loop by

$$B = \frac{\mu_0}{2\pi} \frac{\boldsymbol{p}_\mathrm{m}}{r^3}. \tag{3.20a}$$

Comparing this expression with that for an electric dipole (1.25) one to recognizes the similarity.

For large distances from the loop ($z \gg R$) we can use the approximation $r = \sqrt{(z^2 + R^2)} \approx z$ and the field on the axis is then

$$\mathrm{B}(\mathrm{z} \gg \mathrm{R}) = \mu_0 \mathrm{p_m}/(2\pi \mathrm{z}^3) \tag{3.20b}$$

For points outside the symmetry axis the calculation of the magnetic field becomes more difficult. For points in the plane of the loop one gets elliptic integrals which can be solved only numerically (see for instance [2–4]). The magnetic field strength in the plane of the loop is shown in Fig. 3.15b as a function of the distance r from the loop center.

3.2.5.3 Helmholtz Coils

Helmholtz coils consist of a pair of parallel coils which are separated by a distance $d = R$ which equals the radius R of the coils (Fig. 3.16). The current through the coils has in both coils the same direction.

At first we regard a setup with arbitrary distance d between the coils. The origin of our coordinate system is in the center of the coil pair. On the symmetry axis (z-axis). The amount of the magnetic field is according to the last section

$$\begin{aligned} B(z) &= B_1\left(z + \frac{d}{2}\right) + B_2\left(z - \frac{d}{2}\right) \\ &= \frac{\mu_0 \cdot I \cdot R^2}{2} \cdot \left\{ \frac{1}{[(z + d/2)^2 + R^2]^{3/2}} + \frac{1}{[(z - d/2)^2 + R^2]^{3/2}} \right\}. \end{aligned} \tag{3.22a}$$

Expanding this expression into a Taylor series around $z = 0$ all terms with odd exponents vanish. This is obvious, because the amount B is symmetrical around $z = 0$. After some tedious calculations we obtain

$$\begin{aligned} B(z) = &\frac{\mu_0 I R^2}{[(d/2)^2 + R^2]^{3/2}} \\ &\cdot \left[1 + \frac{3}{2} \frac{d^2 - R^2}{(d^2/4 + R^2)^2} z^2 + \frac{15}{8} \frac{(d^4/2) - 3d^2R^2 + R^4}{(d^2/4 + R^2)^4} z^4 + \cdots \right]. \end{aligned} \tag{3.22b}$$

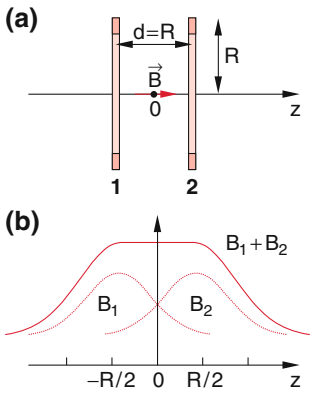


Fig. 3.16 Magnetic field of a Helmholtz coil pair **a**) experimental arrangement, **b**) magnetic field along the z-axis

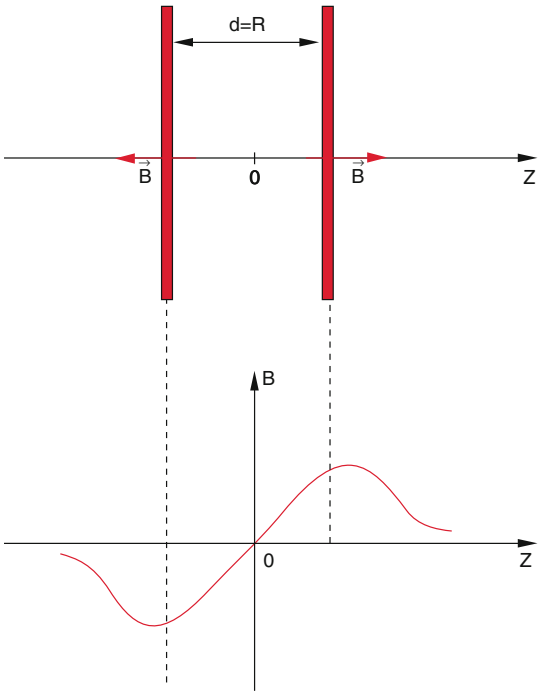


Fig. 3.17 Anti-Helmholtz field with two coils carrying opposite currents

For $d = R$ (Helmholtz condition) the term with z^2 is zero and the field around $z = 0$ is approximately constant. We get:

$$B(z) \approx \frac{\mu_0 I}{(5/4)^{3/2} R} \left[1 - \frac{144}{125} \frac{z^4}{R^4} \right]. \tag{3.22c}$$

For a ratio $z/R = 0.3$ the relative deviation of $B(z)$ from $B(0)$ is less than 1%.

This means that the magnetic field between the two coils is nearly homogeneous.

Three mutually orthogonal Hemholtz pairs of coils are used to compensate external magnetic fields such as the earth magnetic field. This allows experiments at zero magnetic fields.

In "Anti-Helmhotz coils" the current through the two coils flows into opposite directions (Fig. 3.17) and a magnetic field $B(z)$ is generated that is zero for $z = 0$ ($B(0) = 0$) and has a nearly linear slope ($B(z) = a \cdot z$) around $z = 0$. Instead of (3.22a) one obtains

$$\begin{aligned} B(z) &= B_1\left(\frac{d}{2} + z\right) - B_2\left(-\frac{d}{2} + z\right) \\ &= \frac{48}{25 \cdot \sqrt{5}} \frac{\mu_0 I}{R^2} z + \cdots \end{aligned} \tag{3.22d}$$

Such a field is for example used in combination with six orthogonal laser beams to trap cold atoms (see [5]).

3.2.5.4 The Magnetic Field of a Cylindrical Solenoid with Finite Length

In Sect. 3.2.3 it was shown, that inside an *infinity* long solenoid with n windings per m carrying the current I a homogeneous magnetic field

$$B = \mu_0 \cdot n \cdot I$$

exists. We will now study the magnetic field for solenoids with a *finite* length L and in particular the decrease of the field at the ends of the solenoid. We choose the midpoint of the solenoid as zero point of our coordinate system and the symmetry axis as z-axis (Fig. 3.18).

The $n \cdot d\xi$ windings with cross section $A = \pi R^2$ within the length interval $d\xi$ contribute to the magnetic field at the point $P(z)$ the amount

$$dB = \frac{\mu_0 \cdot I \cdot A \cdot n \cdot d\zeta}{2\pi[R^2 + (z-\zeta)^2]^{3/2}}. \tag{3.23}$$

The total field at $P(z)$ is obtained by integration over all windings from $\xi = -L/2$ to $\xi = +L/2$. With the substitution $z - \zeta = R \cdot \tan\alpha$ the integral can be solved and gives

$$B(z) = \int_{-L/2}^{+L/2} dB = -\frac{\mu_0 I \cdot n}{2} \int_{\alpha_1}^{\alpha_2} \cos\alpha \cdot d\alpha$$
$$= \frac{\mu_0 \cdot n \cdot I}{2} \cdot \left\{ \frac{z + L/2}{\sqrt{R^2 + (z+L/2)^2}} - \frac{z - L/2}{\sqrt{R^2 + (z-L/2)^2}} \right\}. \tag{3.24}$$

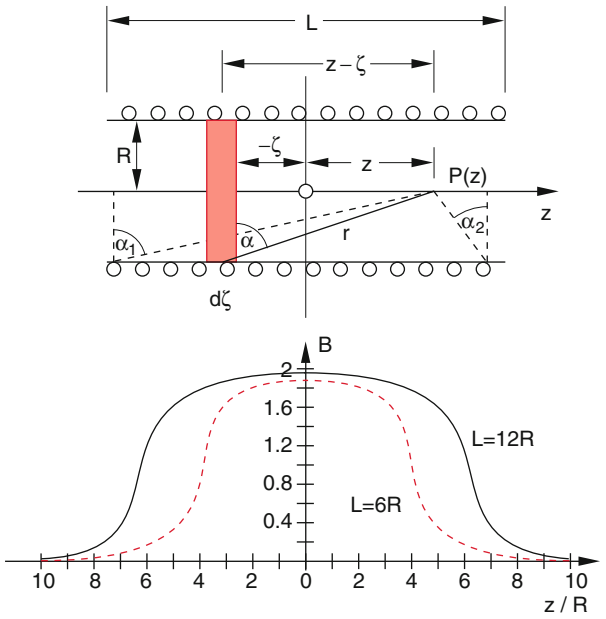


Fig. 3.18 Illustration of the edge effects of the magnetic field of a solenoid

At the midpoint ($z = 0$) of the solenoid the magnetic field becomes

$$B(z=0) = \frac{\mu_0 \cdot n \cdot I}{2} \cdot \frac{L}{\sqrt{R^2 + L^2/4}} \approx \mu_0 \cdot n \cdot I \quad \text{for} \quad L \gg R. \tag{3.25}$$

At the ends of the solenoid at $(z = \pm L/2)$ the field on the symmetry axis has dropped to

$$B(z = \pm L/2) = \frac{\mu_0 \cdot n \cdot I}{2} \cdot \frac{L}{\sqrt{R^2 + L^2}} \approx \mu_0 \cdot \frac{n \cdot I}{2} \quad \text{for} \quad L \gg R \tag{3.26}$$

which is only $1/2$ of its value at $z = 0$.

For reference points far outside the solenoid $(z \gg L \gg R)$ we can expand the radicand in (3.24) in a power series of $R/(z \pm L/2)$ and obtain:

$$B(z) \approx \frac{\mu_0 \cdot n \cdot I \cdot \pi \cdot R^2}{4\pi} \times \left\{ \frac{1}{(z - L/2)^2} - \frac{1}{(z+L/2)^2} \right\}. \tag{3.27}$$

The long solenoid with the cross section $A = \pi \cdot R^2$ acts on far reference points like a magnetic rod with the pole strength

$$p = \pm\mu_0 \cdot n \cdot I \cdot A = B(z=0) \cdot A. \tag{3.28}$$

3.3 Forces on Moving Charges in Magnetic Fields

When charges are moving in magnetic fields an additional force appears besides the Coulomb force between charges. The magnitude and direction of this force can be obtained by some basic experiments:

- A straight wire carrying the current I is freely suspended in the field of a horseshoe magnet (Fig. 3.19). One

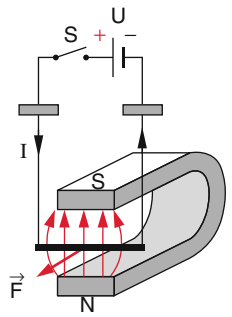


Fig. 3.19 The force on a current carrying conductor in a magnetic field $\boldsymbol{B}$ is perpendicular to $\boldsymbol{B}$ and to the electric current I

observes that the wire is deflected perpendicular to the magnetic field $\boldsymbol{B}$ and to the direction of the current I. The reversal of the current direction or of the magnetic field cause a reversal of the direction of the force.

- When the electric currents I_1 and I_2 flow through two parallel wires (Fig. 3.20) the two wires attract each other, if I_1 and I_2 are flowing into the same direction. They repel each other for opposite current directions. The force between the two wires is proportional to the product $I_1 \cdot I_2$. Since a conductor carrying an electric current I generates a magnetic field and because electric currents are due to moving charges, we conclude that a force acts on moving charges in magnetic fields.
- When the electron beam in a cathode ray tube traverses a magnetic field (Fig. 3.21) the electrons are deflected. Experiments with different directions of the magnetic field prove, that the force acting on the electrons is always perpendicular to the field and to the direction of the velocity. The magnetic field can be generated, for example, by Helmholtz coils (Sect. 3.2.5.3). This allows an easy change of magnitude and direction of the magnetic field just by altering the current through the coils and by turning the coils. The velocity of the electrons can be varied by altering the acceleration voltage.

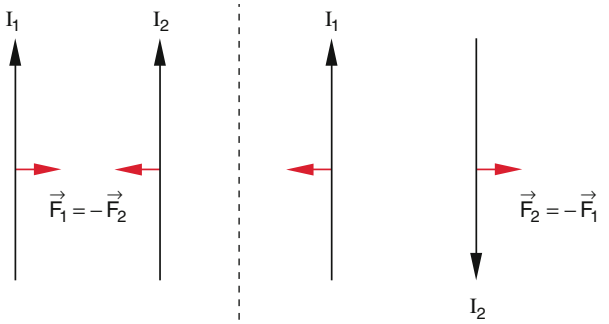


Fig. 3.20 Between two parallel current carrying wires an attractive force acts, if I_1 and I_2 are parallel, whereas a repulsive force acts between anti-parallel currents

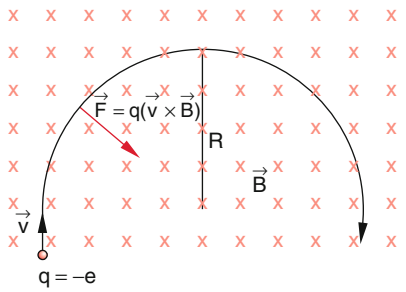


Fig. 3.21 Deflection of an electron beam in a homogeneous magnetic for vertical injection of the beam into the field perpendicular to the drawing plane, inertial systems

The result of such experiments and many more other experiments is: the force $\boldsymbol{F}$ on the electrons with velocity $\boldsymbol{v}$, causing their deviation in the magnetic field $\boldsymbol{B}$, is proportional to the vector product

$$\boldsymbol{F} = k \cdot q \cdot (\boldsymbol{v} \times \boldsymbol{B}),$$

where k is a proportionality factor. In the international system SI the electric current is $I = q \cdot v$ is defined in such a way (Sect. 3.3.1), that the dimensionless constant k becomes $k = 1$, if the force is given in Newton (N) the charge q in As and the velocity in m/s. The magnetic field strength $\boldsymbol{B}$ with $|\boldsymbol{B}| = F/(q \cdot v)$ is therefore determined by the force $\boldsymbol{F}$ on charges q moving with the velocity v: Its unit is then with 1 N m = 1 V A s

$$[B] = 1\,\frac{\mathrm{N}}{\mathrm{As\,m/s}} = 1\,\frac{\mathrm{N}}{\mathrm{A\,m}} = 1\,\frac{\mathrm{V\,s}}{\mathrm{m}^2} = 1\,\mathrm{T},$$

as has been already discussed in Sect. 3.1.

The force

$$\boldsymbol{F} = q \cdot (\boldsymbol{v} \times \boldsymbol{B}). \tag{3.29a}$$

is called the **Lorentz force**. If in addition an electric field $\boldsymbol{E}$ superimposes the magnetic field B, the total force on the charge q is

$$\boldsymbol{F} = q \cdot (\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}). \tag{3.29b}$$

This general Eq. (3.29b) was postulated by *Hendrik Antoon Lorentz (1853–1928)* and is therefore called the general Lorentz force.

We will discuss in Sect. 3.4 the deeper relation between electric and magnetic fields and their mutual connections.

3.3.1 Forces on Conductors with Currents

The electric current I in a conductor with charge density $\varrho = n \cdot q$ and cross section A is according to (2.6a)

$$I = n \cdot q \cdot v_\mathrm{D} \cdot A,$$

when the charges q move with the drift velocity v_D. The Lorentz force on the length dL of the conductor containing $n \cdot A \cdot dL$ charges q is

$$\begin{aligned} \mathrm{d}\boldsymbol{F} &= n \cdot A \cdot \mathrm{d}L \cdot q \cdot (\boldsymbol{v}_\mathrm{D} \times \boldsymbol{B}) \\ &= (\boldsymbol{j} \times \boldsymbol{B}) \cdot \mathrm{d}V, \end{aligned} \tag{3.30a}$$

where $dV = A \cdot dL$ is the volume of the conductor section. The total force on the conductor with length L and current density $\boldsymbol{j} = \boldsymbol{I}/A$ in the magnetic field $\boldsymbol{B}$ is then

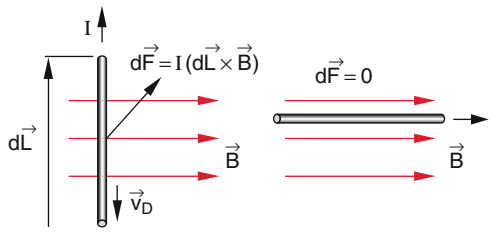


Fig. 3.22 Force onto an electric current perpendicular resp. paralleled to the magnetic field $\boldsymbol{B}$. Note that the drift velocity of the electrons is in the opposite direction as the technically defined direction of I

$$\mathbf{F} = \int (\mathbf{j} \times \mathbf{B}) \mathrm{d}V. \tag{3.30b}$$

Note If the current is caused by electrons (as in all metallic conductors), the charge is $q = -e$ and $\boldsymbol{j}$ points into the opposite direction of v_D, this means that $\boldsymbol{j} \times \boldsymbol{B}$ forms a lefthanded screw.

In case of a straight wire in a homogeneous magnetic field B (Fig. 3.22) $\boldsymbol{j}$ and $\boldsymbol{B}$ are spatially constant. The Lorentz force on the section dL is then

$$\mathrm{d}\mathbf{F} = I \cdot (\mathrm{d}\mathbf{L} \times \mathbf{B}). \tag{3.31}$$

3.3.2 Forces Between Two Parallel Conductors

We will shortly discuss the definition of the unit 1 A of the electric current by the force on two parallel wires carrying the current I (Fig. 3.23). The Lorentz force on a charge $\mathrm{d}q = \rho \cdot A \cdot \mathrm{d}L$ which moves with the drift velocity v_D through the wire 1 with cross section A and length dL in the magnetic field generated by wire 2 is

$$\mathrm{d}\boldsymbol{F} = \mathrm{d}q \cdot (\boldsymbol{v}_\mathrm{D} \times \boldsymbol{B}) = I_1 \cdot (\mathrm{d}\boldsymbol{L} \times \boldsymbol{B}).$$

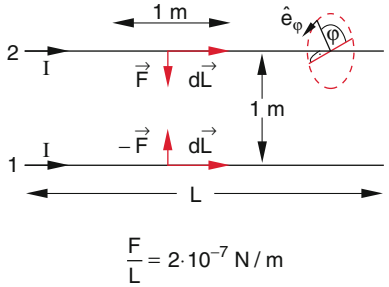


Fig. 3.23 The definition of the unit 1 A of the electric current I

The magnetic field generated by wire 2 is according to (3.8)

$$\boldsymbol{B} = \frac{\mu_0}{2\pi r} \cdot I_2 \cdot \hat{\boldsymbol{e}}_\varphi,$$

where $\hat{e}_\varphi$ is the unit vector in φ-direction (tangent to a circle around wire 2). For parallel wires in z-direction is $\boldsymbol{B} \perp \boldsymbol{v}_\mathrm{D}$. The amount of the force per m of the wire (L = 1 m) is then for a distance $r = R$ between the wires according to (3.31)

$$\frac{F}{L} = I_1 \cdot \frac{\mu_0}{2\pi} \cdot \frac{I_2}{R} = \frac{\mu_0 \cdot I^2}{2\pi R}, \tag{3.32}$$

If the same current I flows through both wires.

For a current I = 1 A the force F/L per m and a distance R = 1 m between the two wires is

$$F/L = \mu_0/2\pi = 2 \times 10^{-7}\ \mathrm{N/m}. \tag{3.33}$$

This equation is used for the definition of the SI-Unit 1 A:

1 A is that electric current, which causes a force $F/L = 2 \times 10^{-7}$ N/m between two parallel wires in vacuum with a distance R = 1 m.

This determines the magnetic constant μ_0 (permeability constant in vacuum) to the exact value

$$\mu_0 = 4\pi \times 10^{-7}\ \mathrm{V\,s/(A\,m)}$$

3.3.3 Experimental Demonstration of the Lorentz Force

The Lorentz force can be quantitatively demonstrated with a focused electron beam in a cell at low gas pressure, which is placed in a homogeneous magnetic field provided by Helmholtz coils (Fig. 3.24). The electrons are bent by the magnetic field into a circular path, which can be seen, because they collide with rest gas atoms, excite them and the excited atoms

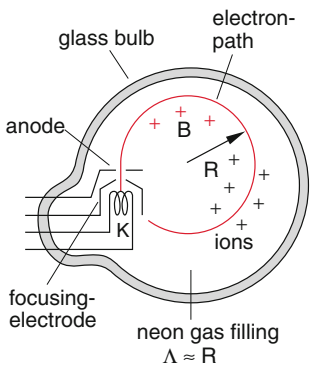


Fig. 3.24 Fadenstrahlrohr

emit visible fluorescence. The electrons are emitted by a hot cathode and accelerated by the voltage U. Their kinetic energy is then $(m/2)v^2 = e \cdot U$. Their velocity is

$$v = \sqrt{\frac{2e \cdot U}{m}}, \tag{3.34}$$

with the initial velocity $\boldsymbol{v}_0 = \{v_x, 0, 0\}$, perpendicular to the magnetic field $\boldsymbol{B} = \{0, 0, B_z\}$. Since the Lorentz force lies, according to (3.29a, 3.29b), in the x-y-plane and is always perpendicular to $\boldsymbol{v}$, the path of the electrons is a circle in the x-y-plane. The Lorentz force acts as centripetal force and we obtain from the condition Lorentz force = centripetal force the equation

$$e \cdot v \cdot B_z = \frac{m \cdot v^2}{R}$$

which allows the determination of the Radius of the circle

$$R = \frac{1}{B} \cdot \sqrt{2m \cdot U/e}. \tag{3.35}$$

The circle can be readily seen by the fluorescence induced by collisions of the electrons with residual gas atoms. Therefore its radius can be measured. The collimated electron beam is not spread out by the collision because of the following facts:

- The density of the atoms is chosen sufficiently low to make the mean free path length $\Lambda = 1/(n \cdot \sigma)$ (σ = collision cross section) larger than the circumference $2\pi R^2$ of the circular path.
- Besides the excitation the electrons can also ionize the atoms of the residual gas. Since the atoms are much heavier than the electrons they diffuse much slower away from the location of their generation. They form a positively charged tube around the electron path, which focuses the electrons onto the center line of this tube.

From the measured values of R, U and B in (3.35) the ratio e/m of electron charge to electron mass can be determined.

When the electrons are injected into the magnetic field $B = \{0, 0, B_z = B\}$ with the velocity $v = \{v_x, v_y, v_z\}$ (Fig. 3.25), the equation of motion

$$m \cdot a = q \cdot (v \times B) \tag{3.36}$$

can be written as the three equations for the components

$$\begin{aligned} m \cdot \dot{v}_x &= -e \cdot v_y \cdot B; \\ m \cdot \dot{v}_y &= +e \cdot v_x \cdot B; \\ m \cdot \dot{v}_z &= 0. \end{aligned}$$

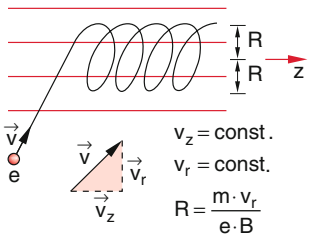


Fig. 3.25 Helical path of electrons injected inclined into a homogeneous magnetic field

The solution gives a helix as trajectory of the electrons in the magnetic field with the radius of the envelope cylinder

$$R = \frac{1}{B} \cdot \sqrt{2m \cdot U/e}$$

and the distance Δz which the electrons move during one circulation period

$$\Delta t = \frac{2\pi \cdot R}{\sqrt{v_x^2 + v_y^2}} = \frac{2\pi \cdot m}{e \cdot B} \tag{3.37a}$$

into the z-direction

$$\Delta z = v_z \cdot \Delta t = \frac{2\pi \cdot m}{e \cdot B} \cdot v_z. \tag{3.37b}$$

Such spiral paths can be made visible with the electron beam tube described above, when the tube is turned in order to change the direction of the injection velocity against the magnetic field.

3.3.4 Electron- and Ion-Optics with Magnetic Fields

The Lorentz force enables the realization of optical systems in magnetic fields, which have found diverse applications as will be illustrated by some examples [3.4].

3.3.4.1 Focussing in Axial Magnetic Fields

The electrons emitted from a hot cathode are accelerated by a voltage U and focused by a special electric field (e.g. an electrically charged hollow cylinder) onto a pinhole aperture at the position $(x = 0,\ y = 0,\ z = 0)$. They leave the pinhole as divergent beam with the velocity $\boldsymbol{v} = \{v_x, v_y, v_z\}$ (Fig. 3.26).

In the axial magnetic field $\boldsymbol{B} = \{0, 0, B_z\}$ they fly on helical trajectories and are focussed again on the z-axis according to (3.37a) after the time

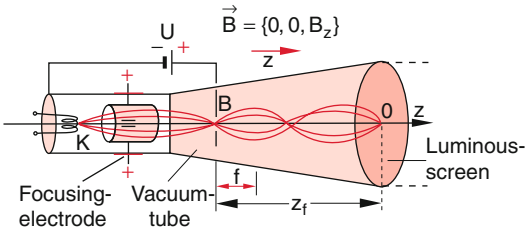


Fig. 3.26 Focusing of electrons in a homogeneous magnetic longitudinal magnetic field, which acts like a lens with the focal length $f = z_f/4$

$$\Delta t = \frac{2\pi \cdot m}{e \cdot B}$$

At $z_f = v_z \cdot \Delta t$. The position z_f is independent of the transverse velocities v_x and v_y. If $v_z \gg \sqrt{v_x^2 + v_y^2}$ we can approximate

$$v_z \approx v = \sqrt{2e \cdot U/m}.$$

The focal length $f = z_f/4$ of this magnetic electron lens is

$$f = \frac{\pi}{B}\sqrt{\frac{m \cdot U}{2e}} \tag{3.38}$$

because a point (i.e. the entrance pinhole) at a distance $2f$ from the symmetry plane at $z = z_f/2 = 2f$ is imaged into a point at $z = z_f = 4f$.

3.3.4.2 Wien-Filter

When an electron beam or an ion beam is sent in the z-direction through a homogeneous magnetic field $\boldsymbol{B} = \{0, 0, B_z\}$ which is superimposed by an electric field $E = \{E_x, 0, 0\}$ (Fig. 3.27) the Lorentz force becomes

$$F = q \cdot (E + v \times B) = q \cdot (E_x - v_z \cdot B_y)\hat{\boldsymbol{e}}_x. \tag{3.39}$$

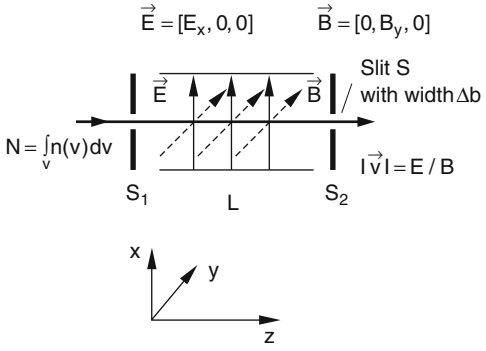


Fig. 3.27 Wien-filter

For the velocity $v_z = E/B = E_x/B_y$ the force is $|F| = 0$. This means, particles in the velocity interval Δv around $v = E/B$ are not or only very slightly deflected and can pass the slit S_2 in Fig. 3.27. Behind the slit one obtains particles with a wanted velocity, which can be selected by choosing the correct values of E and B. The width Δv of the transmitted velocity interval depends on the slit width Δb, the pathlength $\Delta z = L$ through the field region and the velocity v. The calculation (see Problem 3.9) gives

$$\Delta v = \frac{2E_{\text{kin}}}{q \cdot L^2 \cdot B} \cdot \Delta b. \tag{3.40}$$

The design which is called *Wien-filter* after its inventor Max C. W. Wien (1866–1938) acts as a velocity selector for electrons and ions.

3.3.4.3 Focussing by a Homogeneous Transverse Magnetic Field

When ions with mass m and charge $q > 0$ enter through a slit S divergent into a magnetic field B (B is perpendicular to the drawing plane in Fig. 3.28) their trajectories are circular paths with radius

$$R = \frac{m \cdot v}{q \cdot B}.$$

An ion with initial velocity v_0 in the drawing plane perpendicular to the line $\overline{SA}$ reaches the point A after traversing the field in a half circle.

The trajectory of another ion with an initial velocity inclined by an angle α against the vertical direction intersects the trajectory of the first ion in the point C and arrives at the point B on the line $\overline{SA}$. For small angles α the distance $\overline{AB}$ is approximately

$$AB \approx 2R \cdot (1 - \cos\alpha) \approx R \cdot \alpha^2. \tag{3.41}$$

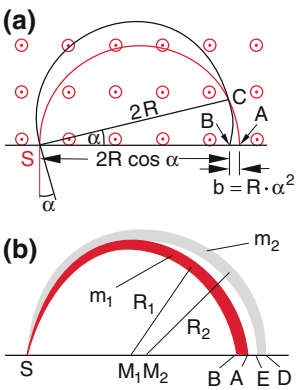


Fig. 3.28 Magnetic sector field as mass selector. **a**) Angular focusing, **b**) mass selection

All ions which exit the slit S within the angular interval $(90° \pm \alpha/2)$ against the line SA are transmitted through an exit slit at the point A with the width $b \approx R \cdot \alpha^2$. This demonstrates that the 180° magnetic sector field images the entrance slit S onto the slit with width AB.

If the source emits ions with different masses m_i within the angular range $90° \pm \alpha/2$, these masses traverse the magnetic field on circular trajectories with different radii

$$R_i = m_i \cdot v_i/(e \cdot B) \tag{3.41a}$$

They arrive therefore at different locations on the line SA. Two masses m_1 and m_2 can be still separated if the arrival interval AB of m_1 does not overlap with the arrival interval DE of m_2. This demands that

$$R_1 - R_2 \geq \frac{1}{4} \cdot R \cdot \alpha^2, \tag{3.41b}$$

where $R = (R_1 + R_2)/2$ is the mean radius.

When the ions are accelerated by a voltage U before they enter the magnetic field, their velocities are

$$v_i = \sqrt{2e \cdot U/m_i}$$

and their radii in the magnetic field are

$$R_i = \frac{1}{B} \cdot \sqrt{2m_i \cdot U/e}. \tag{3.41c}$$

The relative mass resolution is then

$$\begin{aligned} \frac{\Delta m}{m} &= \frac{R_1^2 - R_2^2}{R^2} \\ &= \frac{(R_1 - R_2) \cdot 2R}{R^2} \geq \alpha^2/2. \end{aligned} \tag{3.42}$$

This shows that the mass resolution does not depend on the radius R but quadratically on the divergence angel α of the initial velocities v_{0i} at the entrance into the magnetic field [6–8].

Example

$$\alpha = 2° \triangleq 0.035\,\text{rad} \Rightarrow \frac{\Delta m}{m} \geq 6.1 \cdot 10^{-4}$$

Two masses m_1 = 1500 AMU and m_2 = 1501 AMU can be still separated.

3.3.5 Hall Effect

If a conductor is placed in a magnetic field the Lorentz force causes a deflection of charged particles in the conductor perpendicular to the magnetic field and to the direction of the current (Fig. 3.29). If the magnetic field is sufficiently weak the charged particles are only slightly deflected. This deflection leads to a partial separation of the positive charge of the ions and the negative charge of the electrons and causes an electric field $\boldsymbol{E}_\text{H}$. The charge separation increases until the force $\boldsymbol{F}_\text{C} = n \cdot q \cdot \boldsymbol{E}_\text{H}$ on the charge carriers due to the electric field just compensates the opposite Lorentz force $\boldsymbol{F}_\text{L} = n \cdot q \cdot (\boldsymbol{v} \times \boldsymbol{B})$. Here n is the number of charges q per m^3.

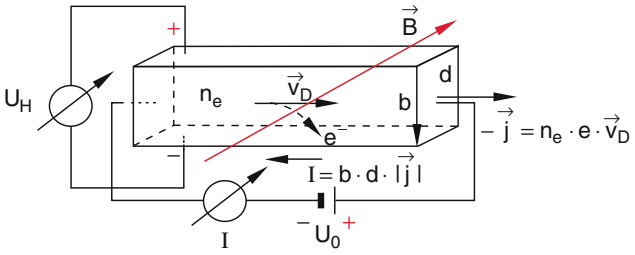


Fig. 3.29 Hall effect

For a conductor with rectangular cross section $A = b \cdot d$ the electric field results in the Hall-voltage

$$U_\text{H} = \int \mathbf{E}_\text{H} \cdot \mathrm{d}\boldsymbol{s} = \boldsymbol{b} \cdot \boldsymbol{E}_\text{H}$$

between the opposing side surfaces with the distance b. In Fig. 3.29 the Hall voltage U_H is the voltage between upper and lower side surface of the conductor. The vector $\boldsymbol{b}$ points therefore downwards. From the relation

$$q \cdot \boldsymbol{E}_\text{H} = -q \cdot (\boldsymbol{v} \times \boldsymbol{B})$$

we obtain with the current density $\boldsymbol{j} = n \cdot q \cdot \boldsymbol{v}$ the Hall voltage

$$U_\text{H} = -\frac{(\boldsymbol{j} \times \boldsymbol{B}) \cdot \boldsymbol{b}}{n \cdot q}. \tag{3.43a}$$

The vector product $\boldsymbol{j} \times \boldsymbol{B}$ points in Fig. 3.29 downwards into the direction of $\boldsymbol{b}$ independent of the sign of the charges, because for positive charges the sign of q and the direction of $\boldsymbol{j}$ both change compared to negative charges. We can therefore write the scalar equation

$$U_\text{H} = -\frac{j \cdot B \cdot b}{n \cdot q} = -\frac{I \cdot B}{n \cdot q \cdot d}. \tag{3.43b}$$

In metals and in most semiconductors the current is supplied by electrons with the charge $q = -e$. This results in a positive Hall voltage

$$U_\text{H} = \frac{I \cdot B}{n \cdot e \cdot d} \tag{3.43c}$$

Some semiconductors show a negative Hall voltage. This can be explained as follows:

In these doped semiconductors mainly electron defects (holes) in the valence band contribute to the current. Such an electron defect acts like a positive charge. If the concentration and drift velocity of the holes are larger than those of the electrons the holes contribute more to the electric current than the electrons.

Measurements of the Hall voltage U_H is a sensitive method fort the determination of weak magnetic fields. Special Hall probes have been developed, which have a very high sensitivity $S = U_H/B$.

For a given current density j the Hall voltage increases with decreasing charge density n! This surprising result is due to the fact that with decreasing values of n the drift velocity v_D increases and with it the Lorentz force. The charge density n is in semiconductors about 10^6-times smaller than in metals. Therefore mainly semiconductors are used as Hall probes [9].

Example

In a Hall probe with $b = 1\,\text{cm}$, $d = 0.1\,\text{cm}$, $n = 10^{15}/\text{cm}^3$ the current density j at a total current $I = 0.1$ A is $j = 1$ A/cm^2. With $e = 1.6 \times 10^{-19}$ C the sensitivity of the Hall probe becomes $S = U_H/B \approx 0.6$ V/T.

For very small magnetic fields a voltage amplifier is necessary which enables the measurements of voltages in the nanovolt range. This allows the determination of magnetic field down to $B < 10^{-6}$ T.

3.3.6 Barlow's Wheel for the Demonstration of "Electron Friction" in Metals

The lower part of a circular aluminum disc emerges into liquid mercury (Fig. 3.30). If a voltage is applied between the axis of the wheel and the mercury trough an electric current flows in radial direction through the disc. If now a magnetic field is applied in axial direction, the electrons in the disc are deflected perpendicular to the current direction, which means the Lorentz force acts in tangential direction. Due to the friction between electrons and metal atoms the whole wheel is starting to rotate around its axis. Reversing the directions of the magnetic field or of the electric current inverts also the direction of the rotation.

This experiment represents a nice demonstration for the model of electric conduction in solids, discussed in Sect. 2.2,

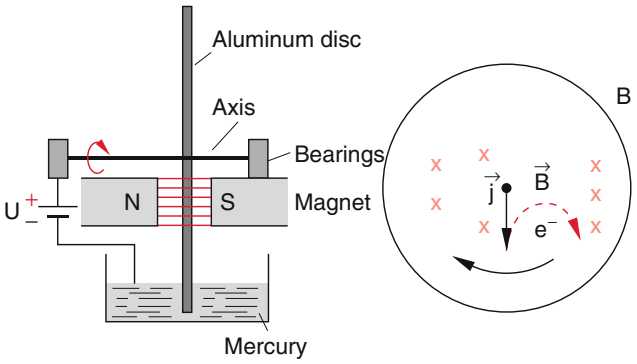


Fig. 3.30 Barlow' s wheel

where the electric resistance is "explained" by the "friction force" between electrons and lattice atoms in a solid conductor.

3.4 Electromagnetic Fields and the Relativity Principle

In Sect. 3.3 the Lorentz force was introduced as an additional force that acts on charges moving in a magnetic field, besides the Coulomb force between charges at rest. We will now show, that the Lorentz force is by no means a principally new force, because it can be directly related to the Coulomb force, if the relativity theory is applied. It should become clear that the relativistic treatment of the Coulomb Law applied to moving charges automatically generates the Lorentz force. This can be realized by the following vivid discussion.

A charge Q resting in an inertial system S' (Fig. 3.31) generates in this system a Coulomb field E'. In another system S, moving against S' with the velocity $\boldsymbol{v}$ the charge Q has the velocity $-\boldsymbol{v}$ and represents therefore for the observer O a current $\boldsymbol{I} = -Q \cdot |\boldsymbol{v}|$ antiparallel to the velocity $\boldsymbol{v}$ of the system S. This current generates a magnetic field $\boldsymbol{B}$ in addition to the electric field $\boldsymbol{E}$.

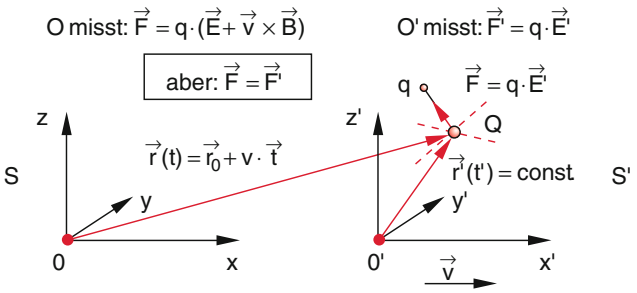


Fig. 3.31 Equivalent description of the force F onto a test charge q in two different

On the other hand all inertial systems are equivalent, i.e. the description of physical laws must be the same for all inertial systems (see Vol. 1 Sect. 3.6). This includes that the forces on a test charge q must be equal for both observers in order to deduce the same equation of motion. When the observer O' in the system S' describes his observations in the system S he has to use the Lorentz transformations and will then arrive at the same laws as the observer O in his system S. Therefore there must be a relation between $\boldsymbol{E}$, $\boldsymbol{E}'$ and $\boldsymbol{B}$ in such a way, that the equivalence of all inertial systems is fulfilled. This implies that the force on the test charge by the fields $\boldsymbol{E}$ and $\boldsymbol{B}$, measured by O in S must be the same as the force caused by $\boldsymbol{E}'$ measured by O'.

This will be discussed in the following more quantitatively. The basic postulates and transformation laws of special relativity, presented in Vol. 1, Chap. 3, are here presupposed [10, 11].

3.4.1 The Electric Field of a Moving Charge

We regard a test charge q which rests in the system S at the point $\{x, y, z\}$, while a field charge Q rests at the origin of the system S', and therefore moves with the velocity $\boldsymbol{v} = (\boldsymbol{v}_x, \boldsymbol{v}_y, \boldsymbol{v}_z) = \{v_x, 0, 0\}$ against S (Fig. 3.32). At the time $t = 0$ the origins of both systems should coincide. We will now calculate the force $\boldsymbol{F} = q \cdot \boldsymbol{E}$ on the test charge q at the time $t = 0$. The electric field is determined by the field charge Q which is moving for the observer O.

The magnitude of the charges Q and q will not be changed by their motion. In the system S the charges have the spatial and time coordinates at time $t = 0$ for $Q = \{0, 0, 0, 0\}$ and for $q = \{x, y, z, 0\}$.

In the system S' where the charge Q rests at the origin and which moves with the velocity $\boldsymbol{v} = \{\boldsymbol{v}_x, 0, 0\}$ against S the charge Q has the spatial-time coordinates $O' = \{0, 0, 0, t'\}$, while the coordinates of q are $\{x', y', z', t'\}$. The Lorentz transformations for length, time, velocity and force between the systems S and S' moving with the velocity $\boldsymbol{v} = (v_x, 0, 0)$

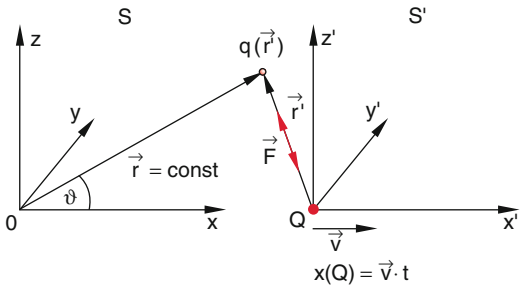


Fig. 3.32 Illustration for the derivation of Eq. (3.45a, 3.45b)

Table 3.1 Lorentz-transformations for lengths, time, velocities and forces

Length and time	Velocities
$x' = \gamma\,(x - v \cdot t)$	$u'_x = \delta\,(u_x - v)$
$y' = y;\quad z' = z$	$u'_y = \frac{\delta}{\gamma}\,u_y$
$t' = \gamma\,(t - \frac{v \cdot x}{c^2})$	$u'_z = \frac{\delta}{\gamma}\,u_z$
where the abbreviations are	
$\gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2}$	$\delta' = \left(1 + \frac{v u'_x}{c^2}\right)^{-1}$
$\delta = \left(1 - \frac{v \cdot u_x}{c^2}\right)^{-1}$	
Forces: $\mathbf{F} = \mathbf{F}'$	
$F'_x = \delta \cdot \left(F_x - \frac{v}{c^2}\,\mathbf{F} \cdot \mathbf{u}\right)$	$F_x = \delta'\left(F'_x + \frac{v}{c^2}\,\mathbf{F} \cdot \mathbf{u}'\right)$
$F_y = \frac{\gamma}{\delta'}\,F'_y\,;\quad F_z = \frac{\gamma}{\delta'}\,F'_z$	$F'_y = \frac{\delta}{\gamma}\,F_y\,;\quad F'_z = \frac{\delta}{\gamma}\,F_z$

against each other are compiled in Table 3.1 in order to remind the reader to the more extensive treatment in Vol. 1, Sect. 3.6 where these formulas had been derived. In our case with $v_x = v$, they reduce to

$$x' = \gamma(x - v \cdot t);\quad y' = y;\quad z' = z;$$
$$t' = \gamma\left(t - \frac{v \cdot x}{c^2}\right).$$

Note that the events for $\{0, 0, 0, 0\}$ and for $\{x, y, z, 0\}$, which occur simultaneously at the time $t = 0$ for the observer O are no longer simultaneous for the observer O'. For him they are for $Q = \{0, 0, 0, 0\}$ and for $q = \{x', y', z', t' = -\gamma \cdot v \cdot x/c^2\}$. In order to determine the force $\boldsymbol{F}$ between Q and q we must know the distance between Q and q. This demands the simultaneous measurements of the coordinates of both charges. Since the field charge Q rests in S' at the origin $0'$ its coordinates remain the same for the times $t' = 0$ and $t' = -\gamma \cdot v \cdot x/c^2$. We can therefore determine the distance $r' = (x'^2 + y'^2 + z'^2)^{1/2}$ unambiguously.

Experiments show, that for field charges Q at rest the Coulomb force between Q and q does not depend on the velocity u of q. as long as v is sufficiently small. The observer O' measures the force

$$\mathbf{F}' = \frac{q \cdot Q}{4\pi \cdot \varepsilon_0} \cdot \frac{\hat{\boldsymbol{r}}'}{r'^2}. \tag{3.44}$$

When we now transform the force components F_x, F_y and F_z according to the Lorentz transformation in Table 3.1 into the system S, we obtain for $u' = 0$ (the field charge Q rests in S')

$$\begin{aligned} F_x &= F'_x = \frac{q \cdot Q \cdot x'}{4\pi \cdot \varepsilon_0 \cdot r'^3}; \\ F_y &= \gamma \cdot F'_y = \frac{\gamma \cdot q \cdot Q \cdot y'}{4\pi \cdot \varepsilon_0 \cdot r'^3}; \\ F_z &= \gamma \cdot F'_z = \frac{\gamma \cdot q \cdot Q \cdot z'}{4\pi \cdot \varepsilon_0 \cdot r'^3}. \end{aligned} \tag{3.45a}$$

Since for $t = 0$ the relations hold

$$x' = \gamma \cdot x; \quad y' = y; \quad z' = z$$
$$\Rightarrow r' = (\gamma^2 \cdot x^2 + y^2 + z^2)^{1/2},$$

we get for the force $\boldsymbol{F}$ the vector equation

$$\boldsymbol{F}(\gamma, \boldsymbol{r}) = \frac{q \cdot Q}{4\pi \cdot \varepsilon_0} \cdot \frac{\gamma \cdot \boldsymbol{r}}{(\gamma^2 x^2 + y^2 + z^2)^{3/2}} = q \cdot \boldsymbol{E}(\gamma \cdot \boldsymbol{r}). \tag{3.45b}$$

We can see from Eq. (3.45b) that for the observer O the force always points into the direction of the line between Q and q, but is no longer spherical symmetric. If q is positioned on the x-axis in the direction of motion of Q, we have $y = z = 0$ and $\boldsymbol{F}$ becomes smaller by the factor $1/\gamma^2$. In the direction perpendicular to the velocity of Q is $x = 0$ and $\boldsymbol{F}$ increases by the factor γ.

The field lines of the electric field

$$\boldsymbol{E} = \frac{Q}{4\pi \cdot \varepsilon_0} \frac{\gamma \cdot \boldsymbol{r}}{(\gamma^2 x^2 + y^2 + z^2)^{3/2}} \tag{3.46a}$$

are shown in Fig. 3.33 for three different velocities $v = 0$, $v = 0.5 \cdot c$ and $v = 0.99 \cdot c$. The comparison with the field lines of a charge at rest illustrates the deviation from a spherical symmetric field.

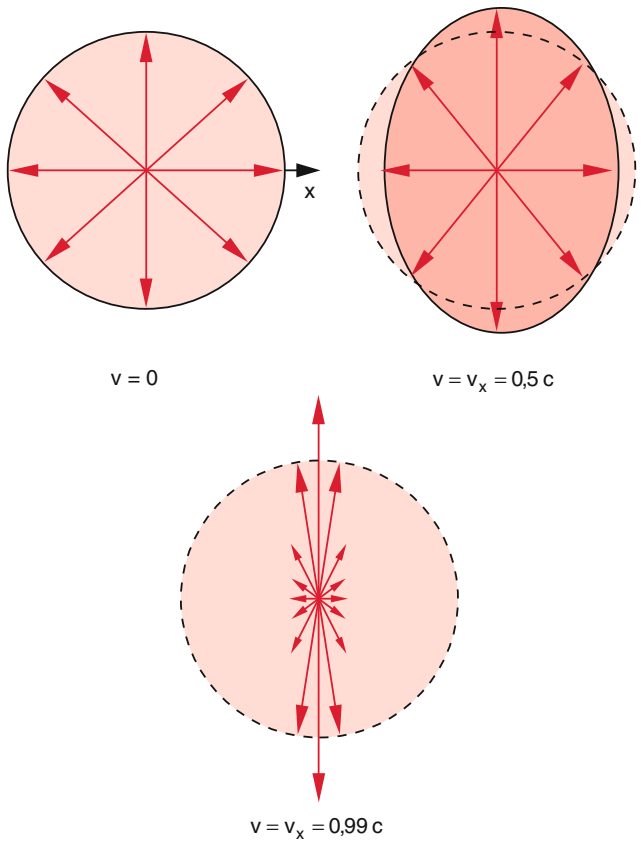


Fig. 3.33 Electric field of moving charge Q for $v = 0$, $v = 0.5 \cdot c$ and $v = 0.99 \cdot c$

Inserting the angles ϑ between the directions of $\boldsymbol{v}$ and of the radius vector $\boldsymbol{r}$ in Fig. 3.32 we can rewrite Eq. (3.46a) using the relations

$$x^2 = r^2 \cdot \cos^2 \vartheta \text{ and } y^2 + z^2 = r^2 \cdot \sin^2 \vartheta.$$

This gives

$$\boldsymbol{E} = \frac{Q}{4\pi \cdot \varepsilon_0 \cdot r^3} \frac{(1 - v^2/c^2) \cdot \boldsymbol{r}}{[1 - (v^2/c^2) \sin^2 \vartheta]^{3/2}}. \tag{3.46b}$$

The result of the above is:

> The electric field of a moving charge is no longer spherical symmetric. The electric field strength $\boldsymbol{E}$ depends on the angle ϑ against the velocity of the charge. It has its maximum for $\vartheta = 90^\circ$, perpendicular to the direction of the velocity v.

3.4.2 Relation Between Electric and Magnetic Field

We now consider the case where both charges $q(0, y, z, t = 0)$ and $Q(0, 0, 0, t = 0)$ move in the system S with the velocity $\boldsymbol{v} = \{v_x, 0, 0\}$ parallel to each other with the constant distance $r = (y^2 + z^2)^{1/2}$ (Fig. 3.34).

In the system S', which moves with the velocity $\boldsymbol{v}$ against S, both charges are at rest. An observer O' in S' therefore measures the Coulomb force components

$$F_{x'} = 0;$$
$$F_{y'} = \frac{q \cdot Q \cdot y'}{4\pi \cdot \varepsilon_0 \cdot r'^3}; \tag{3.47}$$
$$F_{z'} = \frac{q \cdot Q \cdot z'}{4\pi \cdot \varepsilon_0 \cdot r'^3}.$$

We now transform these force components into the system S. Since q does not move in S' we set $u' = 0$ in Table 3.1 and obtain in the system S the force components

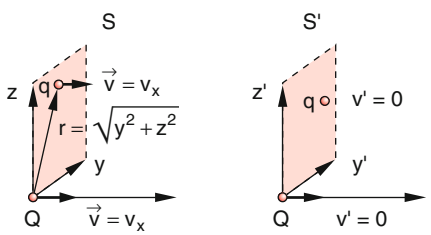


Fig. 3.34 The two charges q and Q rest in the system S' and have therefore the same velocity $v = v_x$ in the system S

$$
\begin{aligned}
F_x &= F'_x = 0; \\
F_y &= \frac{F'_y}{\gamma} = \frac{q \cdot Q \cdot y}{4\pi \cdot \varepsilon_0 \cdot \gamma \cdot r'^3}; \\
F_z &= \frac{F'_z}{\gamma} = \frac{q \cdot Q \cdot z}{4\pi \cdot \varepsilon_0 \cdot \gamma \cdot r'^3}.
\end{aligned}
\tag{3.48}
$$

If q would rest in S we would obtain, according to (3.45a, 3.45b), the force components in S

$$
\begin{aligned}
F_x &= 0; \\
F_y &= \frac{\gamma \cdot q \cdot Q \cdot y}{4\pi \cdot \varepsilon_0 \cdot r'^3}; \\
F_z &= \frac{\gamma \cdot q \cdot Q \cdot z}{4\pi \cdot \varepsilon_0 \cdot r'^3}; \\
\Rightarrow F &= \frac{\gamma \cdot q \cdot Q}{4\pi \cdot \varepsilon_0 \cdot r'^3}\{0, y, z\},
\end{aligned}
\tag{3.45c}
$$

where we have used $y' = y, z' = z \Rightarrow r' = r$

If the description in both systems S and S' should give the same results the difference between (3.48) and (3.45c)

$$
\begin{aligned}
\Delta \boldsymbol{F} &= \frac{q \cdot Q}{4\pi \cdot \varepsilon_0 \cdot r'^3}\left(\frac{1}{\gamma} - \gamma\right)\{0, y, z\} \\
&= \boldsymbol{F}_{\text{magn}}
\end{aligned}
\tag{3.49}
$$

must be caused by the magnetic Lorentz force $\boldsymbol{F}_{\text{magn}} = q \cdot (\boldsymbol{v} \times \boldsymbol{B})$ which the observer O in S postulates, according to (3.29a). Inserting into (3.49) gives

$$
q(\boldsymbol{v} \times \boldsymbol{B}) = -\frac{q \cdot Q}{4\pi \cdot \varepsilon_0 \cdot r'^3} \cdot \gamma \cdot (v^2/c^2) \cdot \{0, y, z\}. \tag{3.50}
$$

The comparison between (3.49) and (3.45c) further demonstrates that between the magnetic force which is measured by O for the field force Q moving with the velocity $\boldsymbol{v}$ and the electric force which would be measured by O for the field charge Q at rest the relation exists

$$
\mathbf{F}_{\text{magn}} = -\left(\frac{v^2}{c^2}\right) \cdot \mathbf{F}_{\text{el}} \,. \tag{3.51}
$$

The additional magnetic force is therefore caused by the motion of the field charge Q. If both charges Q and q would move with the velocity c of light against the system of the observer, we would get

$$
\boldsymbol{F}_{\text{magn}} = -\boldsymbol{F}_{\text{el}} \quad \text{for } v = c
$$

This implies that the total force between the two charges moving parallel to each other with $v = c$ (Fig. 3.35) would be zero. This situation can be indeed approximately realized in particle accelerators (see Vol. 4), where electrons or protons are accelerated to velocities $v \geq 0.99999c$. The repulsive Coulomb forces between the electrons in an electron beam, which would destroy the collimation of the beam

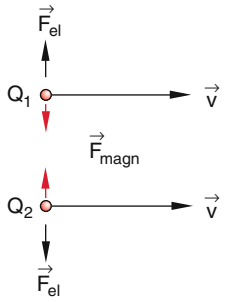


Fig. 3.35 Electric and magnetic forces between two charges Q_1 and Q_2 of equal sign that move with the same velocity

and would spread out the electrons, is nearly compensated by the attractive magnetic force (3.51).

From the magnetic and electric forces

$$
\boldsymbol{F}_{\text{magn}} = q \cdot (\boldsymbol{v} \times \boldsymbol{B}) \quad \text{and} \quad \boldsymbol{F}_{\text{el}} = q \cdot \boldsymbol{E}
$$

acting on a charge q which moves with the velocity v, measured by the observer O in the system S, the relation between electric and magnetic field can be obtained. Inserting into (3.51) gives

$$
\boldsymbol{E} = -\left(\frac{c^2}{v^2}\right) \cdot (\boldsymbol{v} \times \boldsymbol{B}) \tag{3.52a}
$$

The vectorial multiplication with $\boldsymbol{v}$ yields

$$
\boldsymbol{v} \times \boldsymbol{E} = -\frac{c^2}{v^2}\boldsymbol{v} \times (\boldsymbol{v} \times \boldsymbol{B}) = -\frac{c^2}{v^2}\left[(\boldsymbol{v} \cdot \boldsymbol{B})\boldsymbol{v} - v^2 \cdot \boldsymbol{B}\right].
$$

The first term in the bracket is zero, because $\boldsymbol{v} \perp \boldsymbol{B}$ (the magnetic field $\boldsymbol{B}$ generated by a charge moving with the velocity $\boldsymbol{v}$ is always perpendicular to the velocity $\boldsymbol{v}$). We obtain then finally

$$
\boldsymbol{B} = \left(\frac{1}{c^2}\right) \cdot (\boldsymbol{v} \times \boldsymbol{E}) \tag{3.52b}
$$

Since $\boldsymbol{B} \perp \boldsymbol{v}$ we get for the amounts of $\boldsymbol{E}$ and $\boldsymbol{B}$ for a charge moving with the velocity $\boldsymbol{v}$ the relation

$$
|\boldsymbol{B}| = \left(\frac{v}{c^2}\right) \cdot |\boldsymbol{E}|. \tag{3.53a}
$$

When the amount v of the velocity $\boldsymbol{v}$ approaches the velocity c of light, (3.53a) reduces to

$$
B = \left(\frac{1}{c}\right) \cdot E \quad \text{with } B = |\boldsymbol{B}| \text{ and } E = |\boldsymbol{E}| \tag{3.53b}
$$

The magnetic field B of a moving charge q can be explained in the relativistic theory as a change of the electric field. The corresponding change $\boldsymbol{\Delta F}$ of the Coulomb force F on a test charge q just gives the Lorentz force $\boldsymbol{F}_L = q \cdot (\boldsymbol{v} \times \boldsymbol{B})$

3.4.3 Relativistic Transformation of Charge Density and Electric Current

We will illustrate again the real origin of the magnetic field of an electric current by a very instructive example: A test charge q moves with the velocity $\boldsymbol{v}$ parallel to a long straight conductor carrying the current I (Fig. 3.36). According to the considerations described in Sect. 3.4.2 the observer in system S where the conductor is at rest, measures the Lorentz force

$$\boldsymbol{F} = q \cdot (\boldsymbol{v} \times \boldsymbol{B})$$

For O the charge densities (charge per m length) in the conductor are λ_+ for the positive ions and $\lambda_- = -\lambda_+$ for the electrons. This means that the conductor is electrically neutral. The electrons move with the drift velocity $\boldsymbol{v}_D$ against the ions resting in the conductor. The current is then $I = \lambda_- \cdot v_\mathrm{D}$.

For an observer O' who moves with the test charge q, i.e. with the velocity v parallel to the conductor the length of the conductor is shortened due to the Lorentz contraction. He therefore measures the higher charge density

$$\lambda'_+ = \frac{\lambda_+}{\sqrt{1 - v^2/c^2}} = \gamma \cdot \lambda_+ \tag{3.54a}$$

for the ions resting in the conductor, and

$$\lambda'_- = \frac{\lambda_0}{\sqrt{1 - v'^2/c^2}} = \gamma' \cdot \lambda_0 \tag{3.54b}$$

for the electrons which move for O' according to the Lorentz transformations for velocities in Table 3.1 with the velocity

$$\boldsymbol{v}' = \frac{\boldsymbol{v}_\mathrm{D} - \boldsymbol{v}}{1 - \boldsymbol{v}_\mathrm{D}\boldsymbol{v}/c^2}$$

Their charge density would be λ_0 for an observer moving with the electrons, i.e. for whom the electrons are at rest. According to (3.54a) we get the electron charge density for O'

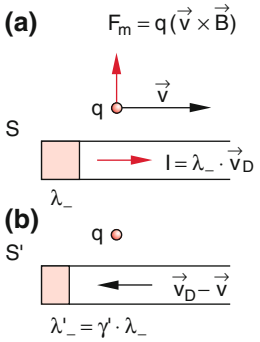


Fig. 3.36 Interaction between a straight conductor with the current I and a charge q which moves with the velocity $v = v_x$ parallel to the wire **a**) in the system S where the conductor is at rest, **b**) in the system S' where the charge q rests and the charges in the wire move with the velocity $v_D = -v$

$$\lambda_- = \frac{\lambda_0}{\sqrt{1 - v_\mathrm{D}^2/c^2}} \,. \tag{3.54c}$$

inserting in (3.54b) gives with the abbreviations $\beta' = v'/c$ and $\beta_\mathrm{D} = v_\mathrm{D}/c$

$$\lambda'_- = \frac{\sqrt{1 - \beta_\mathrm{D}^2}}{\sqrt{1 - \beta'^2}} \cdot \lambda_- .$$

Based on the relativistic addition theorem for velocities (see Table 3.1)

$$\beta' = \frac{\beta_\mathrm{D} - \beta}{1 - \beta \cdot \beta_\mathrm{D}}$$

$$(\beta = v/c)$$

we can eliminate β' and obtain finally

$$\begin{aligned} \lambda'_- &= \frac{1 - \beta \cdot \beta_\mathrm{D}}{\sqrt{1 - \beta^2}} \cdot \lambda_- \\ &= \gamma \cdot (1 - \beta \cdot \beta_\mathrm{D}) \cdot \lambda_- . \end{aligned} \tag{3.54d}$$

$$\begin{aligned} \lambda' &= \frac{1}{\sqrt{1 - \beta^2}} \cdot \lambda_+ + \frac{1 - \beta \cdot \beta_D}{\sqrt{1 - \beta^2}} \cdot \lambda_- \\ &= \gamma \cdot (v/c^2) \cdot v_D \cdot \lambda_+ , \end{aligned} \tag{3.55}$$

where we have used $\lambda_+ = -\lambda_-$.

The current I is for the observer O at rest in S

$$I = \lambda_- \cdot v_\mathrm{D}$$

For the moving observer O', however, the current is

$$I' = \lambda'_+ \cdot (-v) + \lambda'_- \cdot v' .$$

Inserting for λ'_+, λ'_- and v' the expressions derived above, and taking into account that $\lambda_+ = -\lambda_-$ one obtains the result

$$I' = \frac{1}{\sqrt{1 - \beta^2}} \cdot I = \gamma \cdot I . \tag{3.56}$$

The moving observer O' measures therefore a larger current $\gamma \cdot I (\gamma > 1)$ than the observer O at rest.

The force on the charge q moving with the velocity v parallel to the conductor is for the observer O' who moves with q

$$\begin{aligned} \boldsymbol{F}' = q \cdot \boldsymbol{E}' &= \frac{q \cdot \lambda' \cdot \hat{\boldsymbol{r}}}{2\pi \cdot \varepsilon_0 \cdot r} \\ &= \gamma \cdot q \cdot (v/c^2) \cdot \frac{I}{2\pi \cdot \varepsilon_0} \cdot \frac{\hat{\boldsymbol{r}}}{r} . \end{aligned} \tag{3.57}$$

The observer O at rest in S measures, according to the Lorentz transformation (Table 3.1) the force

$$\boldsymbol{F} = \boldsymbol{F}'/\gamma = q \cdot v/c^2 \cdot \frac{I \cdot \hat{\boldsymbol{r}}}{2\pi \cdot \varepsilon_0 \cdot r}. \tag{3.58}$$

The magnetic field of a straight conductor (3.17) has the amount

$$B = \mu_0 \cdot I/(2\pi r)$$

and its direction is perpendicular to $\boldsymbol{v}$ and $\boldsymbol{r}$. Therefore (3.58) can be also written as

$$\boldsymbol{F} = q \cdot \frac{1}{c^2 \cdot \varepsilon_0 \cdot \mu_0} \cdot (\boldsymbol{v} \times \boldsymbol{B}). \tag{3.59}$$

This is identical to the Lorentz force (3.29a) if the relation

$$\varepsilon_0 \cdot \mu_0 = 1/c^2 \tag{3.60}$$

between the permittivity constant ε_0, the permeability constant μ_0 and the speed of light c is fullfilled (see Sect. 7.1).

It is remarkable, that the difference in the Lorentz contraction of the conductor length for the ions resting in the conductor and the electrons moving with the small drift velocity v_D is caused by the small drift velocity of some mm/s. The much larger thermal velocity of the electrons, which is randomly distributed into all directions, has the average zero and is therefore unimportant for the Lorentz transformation.

One should, however, point out to the following facts:

For an electrically neutral conductor the electric forces acting on a test charge due to electrons and ions would be completely compensated if the test charge is at rest relative to the conductor. If it moves relative to the conductor the compensation is no longer complete, but there remains a résidual charge

$$\Delta Q = \gamma \cdot (v \cdot v_{\mathrm{D}}/c^2) \cdot Q \tag{3.61}$$

of the total ion charge. The electric force of this residual charge is equal to the magnetic Lorentz force $\boldsymbol{F} = q \cdot (\boldsymbol{v} \times \boldsymbol{B})$.

Summarizing we can say:

The magnetic field of an electric current and the Lorentz force acting on a moving charge can be deduced by the relativity theory from the Coulomb law and the Lorentz transformations. The magnetic field is therefore not a property of charged matter independent of the electric field, but is in fact a consequence of the change in the electric field of moving charges, due to the Lorentz contraction. One therefore speaks of the **electromagnetic field** of moving charges.

3.4.4 Equations for the Transformation of Electromagnetic Fields

We will now consider the magnetic field of an electric current from another point of view. For this purpose we deduce the equations for the transformation of electromagnetic fields ($\boldsymbol{E}$, $\boldsymbol{B}$) for transitions from a system S at rest to a moving system S'. We consider the case that in the lab system S at rest the two charges $Q(x(t), 0, 0)$ and $q(x(t), y, z)$ both move with the velocity $\boldsymbol{v} = \{v_x, 0, 0\}$ parallel to each other at the distance $\mathrm{r} = (\mathrm{y}^2 + \mathrm{z}^2)^{1/2}$. Since the system S' moves with the velocity v against S the two charges both rest in S'. The observer O in S measures the force components

$$\begin{aligned} F_x &= q \cdot E_x; \\ F_y &= q \cdot (E_y - v_x \cdot B_z); \\ F_z &= q \cdot (E_z + v_x \cdot B_y) \end{aligned} \tag{3.62}$$

acting on the test charge q and he concludes the existence of an electric and magnetic field.

The observer O' in S', who moves together with both charges q and Q, measures only an electric field $\boldsymbol{E}'$, which differs, however, from the electric field $\boldsymbol{E}$ measured by O. He measures the force components

$$\begin{aligned} F'_x &= q \cdot E'_x; \\ F'_y &= q \cdot E'_y; \\ F'_z &= q \cdot E'_z. \end{aligned} \tag{3.63}$$

For the transformation of the force components from S to S' the Lorentz transformations (see Table 3.1) must be valid.

Note, that in S' both charges are at rest and therefore is $\boldsymbol{u} = \boldsymbol{0}$.

We therefore get

$$F'_x = F_x; \quad F'_y = \gamma \cdot F_y; \quad F'_z = \gamma \cdot F_z.$$

This gives the relation between the fields $\boldsymbol{E}$, $\boldsymbol{B}$ and $\boldsymbol{E}'$:

$$\begin{aligned} E'_x &= E_x; \\ E'_y &= \gamma \cdot (E_y - v_x \cdot B_z); \\ E'_z &= \gamma \cdot (E_z + v_x \cdot B_y). \end{aligned} \tag{3.64a}$$

The back transformation from S' to S for the case where Q rests in S and therefore moves in S' has the consequence that O' now measures an electric and a magnetic field. With $v'_x = -v_x$ we get

$$\begin{aligned} E_x &= E'_x; \\ E_y &= \gamma \cdot (E'_y + v_x \cdot B'_z); \\ E_z &= \gamma \cdot (E'_z - v_x \cdot B'_y). \end{aligned} \tag{3.64b}$$

For the general case where the charge Q moves in S as well as in S' both observers measures an electric and a magnetic field but of different magnitude. The corresponding transformation equations can be obtained from (3.64a, 3.64b) and from the Lorentz transformations for velocities (Vol. 1, Eq. (3.28). The result is

$$\begin{aligned} B'_x &= B_x; \\ B'_y &= \gamma \cdot \left(B_y + \frac{v}{c^2} \cdot E_z\right); \\ B'_z &= \gamma \cdot \left(B_z - \frac{v}{c^2} \cdot E_y\right), \end{aligned} \tag{3.65a}$$

The associated back transformations yield

$$\begin{aligned} B_x &= B'_x; \\ B_y &= \gamma \cdot \left(B'_y - \frac{v}{c^2} \cdot E'_z\right); \\ B_z &= \gamma \cdot \left(B'_z + \frac{v}{c^2} \cdot E'_y\right). \end{aligned} \tag{3.65b}$$

The Eqs. (3.64a, 3.64b) and (3.65a, 3.65b) couple the electric and the magnetic fields. These coupled fields are therefore called the ***electromagnetic field***. The separation into a pure electric or pure magnetic field depends on the reference system in which the observed process is described.

Note, however, that all observers in arbitrary inertial systems come to the same conclusion for the equations of motion, without any inconsistencies.

3.5 Matter in Magnetic Fields

In this section we will discuss the magnetic phenomena which are observed when matter is brought into an external magnetic field. We will do this in a more phenomenological way because the microscopic model of magnetism, based on an atomic theory can be treated only after the introduction of atomic physics in Vol. 3.

The magnetic phenomena, discussed here are completely equivalent to the corresponding dielectric polarization introduced in Sect. 1.7. We start with the important definition of the magnetic dipole.

3.5.1 Magnetic Dipoles

We have seen in Sect. 3.2.6 that the magnetic field of a plane current loop is equal to that of a short magnetic rod, called a magnetic dipole. We define the product

$$\boldsymbol{P}_\mathrm{m} = I \cdot \boldsymbol{A} \tag{3.66}$$

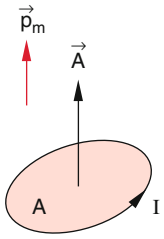


Fig. 3.37 Magnetic dipole moment $\boldsymbol{p}_\mathrm{m}$ of an area A circumvented by the current I

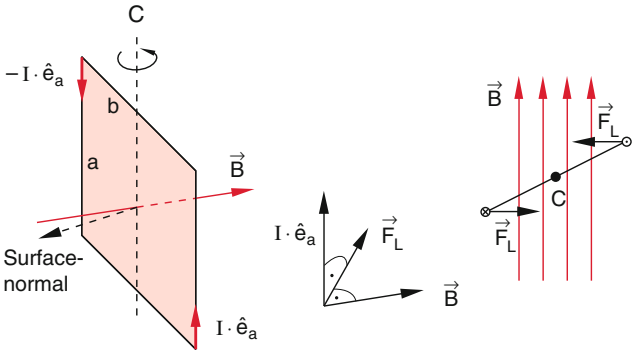


Fig. 3.38 Torque acting onto a rectangular current loop

of current I and area A with surface vector $\boldsymbol{A}$, enclosed by the current loop as the magnetic dipole moment $\boldsymbol{p}_\mathrm{m}$ of the current loop. The direction of $\boldsymbol{p}_\mathrm{m}$ is defined such, that it forms a right hand screw with the current direction (Fig. 3.37).

When the current loop is placed in a magnetic field, the Lorentz forces acting on the dipole cause a torque. We will calculate this for the example of a rectangular loop, which can rotate about an axis C (Fig. 3.38).

The Lorentz force acting on the two sides a of the rectangular loop with the area $A = a \cdot b$ is

$$\boldsymbol{F} = a \cdot I \cdot (\hat{\boldsymbol{e}}_a \times \boldsymbol{B}),$$

where $\hat{\boldsymbol{e}}_a$ is the unit vector in the direction of a and $\boldsymbol{I} \cdot \hat{\boldsymbol{e}}_a$ is the technical current direction (opposite to the drift velocity of the electrons). The forces on the sides b are compensated by the restoring force of the suspension wire.

The forces on the sides a cause a torque

$$\begin{aligned} \boldsymbol{D} &= 2 \cdot \frac{b}{2} \cdot (\hat{\boldsymbol{e}}_b \times \boldsymbol{F}) \\ &= a \cdot b \cdot I \cdot (\hat{\boldsymbol{e}}_b \times \hat{\boldsymbol{e}}_a) \times \boldsymbol{B} = \boldsymbol{I} \cdot \boldsymbol{A} \times \boldsymbol{B}. \end{aligned}$$

Inserting the magnetic dipole moment $\boldsymbol{p}_\mathrm{m} = I \cdot \boldsymbol{A}$ we obtain

$$\boldsymbol{D} = \boldsymbol{p}_\mathrm{m} \times \boldsymbol{B} \tag{3.67}$$

Note the analogy to the electric case where the torque on an electric dipole in an electric field $\boldsymbol{E}$ is $\boldsymbol{D} = \boldsymbol{p}_{el} \times \boldsymbol{E}$.

In the same way the potential energy of the dipole is analogue for both cases (see Sect. 1.4.1). In the magnetic case the potential energy is

$$W = -\boldsymbol{p}_\mathrm{m} \cdot \boldsymbol{B}. \tag{3.68a}$$

And in the electric case

$$\mathrm{W} = -\mathbf{p}_\mathbf{el} \cdot \mathbf{E} \tag{3.68b}$$

The force onto a magnetic dipole in a homogeneous magnetic field is zero! In an inhomogeneous field it is

$$\boldsymbol{F} = \boldsymbol{p}_\mathrm{m} \cdot \mathbf{grad}\, \boldsymbol{B}. \tag{3.69}$$

The Eqs. (3.67)–(3.69) do not contain the specific geometrical form of the current loop. They are therefore valid for any magnetic dipole, e.g. also for permanent magnetic rods.

In the following sections some examples of magnetic dipoles and their applications are given.

3.5.1.1 Moving Coil Instruments

The torque on current coils in magnetic fields is used for the measurement of small currents. A small rectangular coil with N windings is suspended by a thin wire in a radial magnetic field (Fig. 2.28). The torque excreted by the magnetic field on the rotatable coil with N windings

$$\boldsymbol{D} = \boldsymbol{P}_\mathrm{m} \times \boldsymbol{B} = N \cdot \boldsymbol{I} \cdot \boldsymbol{A} \times \boldsymbol{B}$$

with the amount $D = I \cdot N \cdot A \cdot B \cdot \sin\alpha = I \cdot N \cdot A \cdot B$ because the surface normal vector $\boldsymbol{A}$ is always perpendicular to the radial magnetic field.

The twist of the coil is given by the equilibrium condition that the magnetic torque equals the opposite torque caused by the torsion of the wire.

With a small mirror the twist can be projected with a light beam on a scale (mirror galvanometer). More robust instruments use instead of the wire a solid axis which is rotatable suspended in ball bearings. A spiral spring provides the restoring torque. The sensitivity is determined by the strength of the spring and the friction of the ball bearings.

3.5.1.2 Atomic Magnetic Moments

A particle with mass m and charge q, that moves with the velocity v on a circle with radius R represents a circular current

$$I = q \cdot v = q \cdot \boldsymbol{v}/(2\pi R).$$

where v is the number of circulations per second, i.e. the orbital frequency. The magnetic moment of this current is

$$\boldsymbol{p}_m = q \cdot \boldsymbol{v} \cdot \boldsymbol{A} = \frac{1}{2} q \cdot R^2 \cdot \boldsymbol{\omega} \tag{3.70}$$

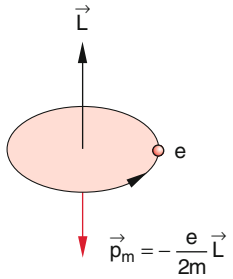


Fig. 3.39 Relation between angular momentum L and magnetic moment pm of a particle with mass m and charge $q = -e$ moving on a circle

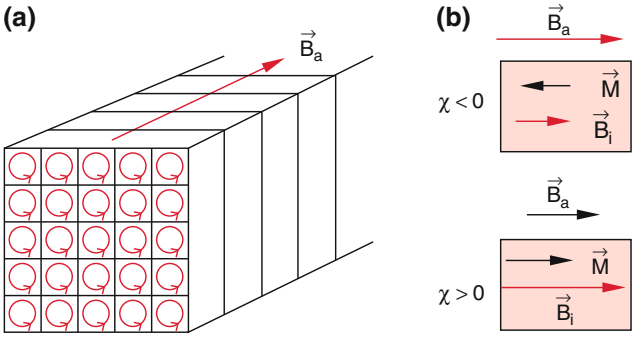


Fig. 3.40 **a**) Magnetization M generated by induced ($\chi < 0$) or permanent ($\chi > 0$) atomic circular currents within the atoms of the magnetic material. Each of these circular currents generates a magnetic dipole moment $\boldsymbol{p}_m$. **b**) The orientation of these dipoles results in the macroscopic magnetization $M = (1/V) \sum \boldsymbol{p}_m$, which either amplify the magnetic field (paramagnetic material) or decrease it (diamagnetic material)

with $\omega = 2\pi v$. The angular momentum of the circulating mass m is

$$\boldsymbol{L} = m \cdot (\boldsymbol{R} \times \boldsymbol{v}) = m \cdot R^2 \cdot \boldsymbol{\omega}. \tag{3.71}$$

We get from (3.70) and (3.71) the relation between angular momentum and magnetic moment of the charged circulating particle (Figs. 3.39 and 3.40)

$$\boldsymbol{p}_\mathrm{m} = \frac{q}{2m} \cdot \boldsymbol{L}. \tag{3.72}$$

Example

In Bohr's atomic model (see Vol. 3) the electron with mass $\mathrm{m_e}$ and charge $q = -e$ moves on a circle about the proton. The amount L of its angular momentum is in units of the Planck constant $\hbar = \mathrm{h}/2\pi$ equal to

$$L = \ell \cdot \hbar \quad \text{with}\, \ell = 1, 2, 3, \ldots$$

The orbital magnetic moment is then

$$\boldsymbol{p}_{\mathrm{m}} = -\left(\frac{e}{2m_{\mathrm{e}}}\right) \cdot \boldsymbol{L} \Rightarrow |\boldsymbol{p}_{\mathrm{m}}| = -l \cdot \left(\frac{e}{2m_e}\right) \cdot \hbar$$

With the orbital magnetic moment for $l = 1 \Rightarrow L = \hbar$. This gives the lowest energy state of the Bohr atom.

remark: The real hydrogen atom has $L = 0$ in its lowest state. This is one of the deficiencies of the Bohr model.

$$\mu_{\mathrm{B}} = \frac{e \cdot \hbar}{2\mu_{\mathrm{e}}} \tag{3.73}$$

is called the ***Bohr magneton***

3.5.2 Magnetization and Magnetic Susceptibility

The magnetic field in vacuum inside a long solenoid with length L, N windings and the winding density $n = N/L$ is according to Sect. 3.2.3 (Fig. 3.41)

$$B_0 = \mu_0 \cdot n \cdot I.$$

Often the magnetic field is characterized by the magnetic intensity

$$\boldsymbol{H} = \boldsymbol{B}/\mu_0$$

If the inside of the solenoid is filled with matter (e.g. iron), one finds that the magnetic flux

$$\Phi_m = \int \boldsymbol{B} \cdot \mathrm{d}A$$

changes by the factor μ. Since the area A did not change, the magnetic field strength B must have been altered to

$$\boldsymbol{B}_{\mathrm{matter}} = \mu \cdot \boldsymbol{B}_{\mathrm{vacuum}} = \mu \cdot \mu_0 \cdot \boldsymbol{H} \tag{3.74}$$

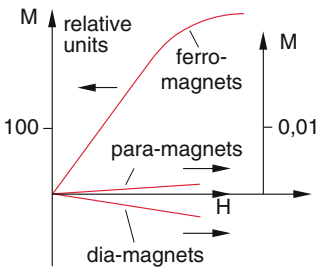


Fig. 3.41 Magnetization $M(H)$ as a function of the magnetic field H for dia- and para-magnetic materials (right side scale) and for ferromagnetic materials (left scale)

The dimensionless constant μ is called the ***relative permeability constant***.

The alteration of the magnetic flux is due to the interaction of the magnetic field with the atoms or molecules in the material. Analogous to the situation in the electric field which induces electric dipoles by charge displacements or orientates already existing **permanent dipoles** and thus generates the dielectric polarisation of matter (see Sect. 1.7) in the magnetic case a macroscopic magnetic polarisation (called **magnetization**) is observed when matter is brought into a magnetic field. It is caused by atomic magnetic moments which are either induced by the magnetic field or oriented in case of permanent magnetic dipoles. The magnetization M is defined as the vector sum

$$\boldsymbol{M} = \frac{1}{V}\sum_{V} \boldsymbol{p}_{\mathrm{m}} \tag{3.75}$$

of the atomic magnetic dipole moments $\boldsymbol{p}_{\mathrm{m}}$ per m^3. The unit of M is

$$|M| = 1\,\frac{\mathrm{A} \cdot \mathrm{m}^2}{\mathrm{m}^3} = 1\,\frac{\mathrm{A}}{\mathrm{m}}$$

The magnetic field strength B in the inside of the solenoid filled with matter is then

$$B = \mu_0 \cdot (H_0 + M) = \mu_0 \cdot \mu \cdot H_0.$$

where $H_0 = H_{\mathrm{vacuum}}$. Experiments show that for not too high fields (see below) the magnetization is proportional to the magnetic intensity H

$$M = \chi \cdot H_0. \tag{3.77}$$

The proportionality factor χ is the **magnetic susceptibility**. Its value decreases generally with increasing temperature.

The comparison of (3.76) and (3.77) show that the following relation exists between χ and the relative permeability μ

$$\begin{aligned} B &= \mu_0 \cdot \mu \cdot H_0 = \mu_0 \cdot (1 + \chi) \cdot H_0 \\ &\Rightarrow \mu = 1 + \chi. \end{aligned} \tag{3.78}$$

Note The constants χ and μ are dimensionless numbers. Often the molar susceptibility χ_{mol} is used with the unit 1/mol. For gases χ_{mol} is the susceptibility of the volume V that is occupied by 1 mol of the gas. Its unit is then m^3/mol.

With respect to their magnetic properties characterized by the value of χ, the magnetic materials are categorized into the following classes:

Table 3.2 Molar magnetic susceptibility χ_{mol} of some dia- and para-magnetic materials and relative permeabilities μ of some ferro-magnets under normal conditions ($p = 10^5$ Pa, T = 0 °C)

Gases	$\chi_{mol}/10^{-12}$ m^3/mol	Material	$\chi_{mol} \cdot 10^9$/mol
(a) *Diamagnetic materials*			
He	−1.9	Cu	−5.46
Ne	−7.2	Ag	−19.5
Ar	−19.5	Au	−28
Kr	−28.8	Pb	−23
Xe	−43.9	Te	−39.5
H_2	−4.0	Bi	−280
N_2	−12.0	H_2O	−13
(b) *Paramagnetic Materials*			
Al	+16.5	O_2	+3450
Na	+16.0	$FeCO_3$	+11,300
Mn(α)	+529	$CoBN_2$	13,000
Ho	72,900	Gd_2O_3	53,200
(c) *Ferro-magnetic Materials*			
Material		μ	
Iron, depending on pre-treatment		500–10,000	
Cobalt		80–200	
Permalloy 78% Ni. 3% Mo		10^4–10^5	
Mu-metal 76% Ni. 5% Cu. 2% Co		10^5	
Supermalloy		10^5–10^6	

Diamagnetic substances: $\chi < 0;\ 10^{-9} < |\chi| < 10^{-6}$
Paramagnetic substances: $\chi > 0;\ 10^{-6} < |\chi| < 10^{-4}$
Ferromagnetic substances: $\chi > 0;\ 10^{2} < |\chi| < 10^{5}$
Anti-ferromagnet: $\chi > 0;\ 0 < |\chi| < 10^{2}$

In Table 3.2 values of χ are compiled for some substances at 0° C temperature [12]. For the molar susceptibility the relation

$$M_{mol} = \chi_{mol} \cdot H_0, \quad \text{with} \quad \chi_{mol} = \chi \cdot V_{mol}$$

between the magnetization M and the magnetic intensity H is equivalent to (3.77), where V_{mol} is the volume occupied by 1 mol of the substance.

3.5.3 Diamagnetism

Diamagnetic substances consist of atoms or molecules that possess no permanent magnetic moment. When such substances are brought into a magnetic field, induced dipoles $\boldsymbol{p}_m$ develop. Their direction is oriented opposite to the inducing magnetic field. This causes a decrease of the field inside the substance, which makes the field inside the sample smaller than outside (see Sect. 4.2).

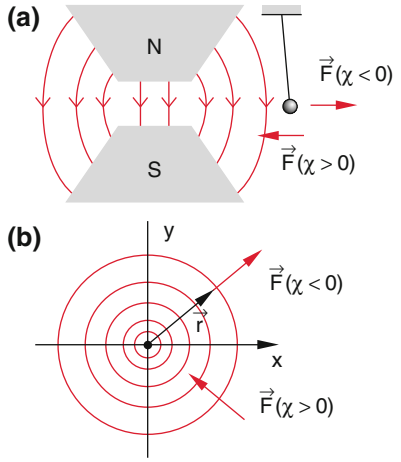


Fig. 3.42 A diamagnetic body in an inhomogeneous magnetic field is pushed out of the high field region. **a**) example of the inhomogeneous field between the poles a an electro-magnet, **b**) inhomogeneous field of a straight wire carrying the current I

The magnetization

$$M = \chi \cdot H \tag{3.79a}$$

Is therefore also opposite to the external field, which implies that χ is negative ($\chi < 0$). The linear relation (3.79a) holds for magnetic fields that are still small compared to the *inner atomic magnetic fields* generated by the motion of the electrons in the shell of the atoms. These field are of the order of 10^2 T or more.

The force on a magnetic dipole in an inhomogeneous magnetic field $\boldsymbol{B}$ is $\boldsymbol{F} = \boldsymbol{p}_m \cdot \operatorname{grad} \boldsymbol{B}$ (see (1.29)).

Since $\boldsymbol{M}$ is antiparallel to $\boldsymbol{B}$ a diamagnetic sample is pushed out of the region of high magnetic fields into ranges of lower field (Fig. 3.42a). The force $\boldsymbol{F}$ onto a sample with volume V and magnetization $\boldsymbol{M} = \chi \cdot \boldsymbol{H} = (\chi/\mu_0) \cdot \boldsymbol{B}$ is

$$\begin{aligned} \boldsymbol{F} &= \boldsymbol{M} \cdot V \cdot \operatorname{grad} \boldsymbol{B} \\ &= (\chi/\mu_0) \cdot V \cdot \boldsymbol{B} \cdot \operatorname{grad} \boldsymbol{B}. \end{aligned} \tag{3.79b}$$

Example

We consider the magnetic field of a straight wire with current I. It is inhomogeneous in the radial direction, since it decreases with 1/r according to (3.17). The magnetic field is

$$\begin{aligned} B &= \frac{\mu_0 I}{2\pi r^2} \cdot \{-y, x, 0\}; \\ \Rightarrow \operatorname{grad} B_x &= \frac{\mu_0 I}{2\pi r^4} \cdot \{2xy, y^2 - x^2, 0\} \\ \operatorname{grad} B_y &= \frac{\mu_0 I}{2\pi r^4} \cdot \{y^2, -x^2 - 2xy, 0\} \end{aligned}$$

With $M = (\chi/\mu_0) \cdot B$ the force on a diamagnetic body with volume V is

$$\begin{aligned} \boldsymbol{F} &= \boldsymbol{M} \cdot V \cdot \mathbf{grad}\, \boldsymbol{B} \\ &= (\chi/\mu_0) \cdot V \cdot \boldsymbol{B} \cdot \mathbf{grad}\, \boldsymbol{B} \end{aligned} \tag{3.79c}$$

Where **grad** $\boldsymbol{B}$ is as the gradient of a vector a tensor and the scalar product $\boldsymbol{B} \cdot \mathbf{grad}\, \boldsymbol{B} = B_x \cdot \mathbf{grad}\, B_x + B_y \cdot \mathbf{grad}\, B_y + B_z \cdot \mathbf{grad}\, B_z$ is a vector with the three components given in the sum above. This finally gives for the force

$$\boldsymbol{F} = \frac{\mu_0 \chi I^2 \cdot V}{4\pi^2 r^4} \cdot \{x, y, 0\} \tag{3.79d}$$

Diamagnetic samples ($\chi < 0$) experience a force into the radial direction away from the wire where the field is weaker, whereas paramagnetic substances and in particular ferromagnetic materials are attracted into the region of higher magnetic field (Fig. 3.42b).

The inhomogeneous magnetic field is often realized by a conical form of the poleshoes in an electromagnet.

The force $\boldsymbol{F}$ acting on magnetic substances in inhomogeneous fields can be used to measure the susceptibility χ by a weighing technique. In the ***Faraday method*** (Fig. 3.43a) the sample which is suspended by a spring is brought into the inhomogeneous magnetic field. Here the force

$$\mathbf{F} = (\chi/\mu_0) \cdot \mathrm{V} \cdot \mathbf{B} \cdot \mathbf{grad}\, \mathbf{B}$$

acts on the sample with volume V, which can be measured by the elongation ($\chi > 0$) or shortening ($\chi < 0$) of the spring. The form of the conical poleshoes is chosen in such a way, that a constant gradient B is achieved over the volume where the sample immerses into the field.

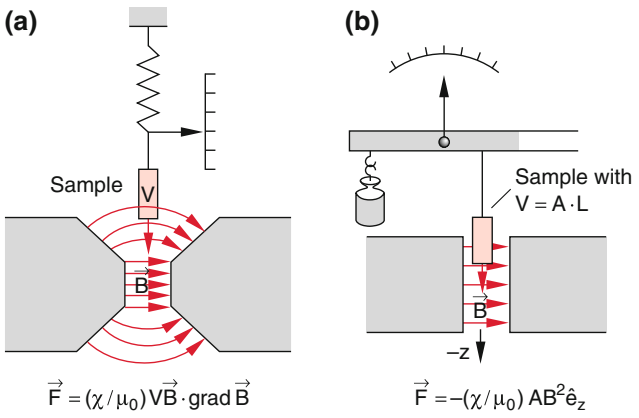


Fig. 3.43 Measurement of the magnetic susceptibly with a balance technique. **a**) Faraday method. **b**) Technique according to Gouy

For the *method of Gouy* the sample immerses only partly into a homogeneous field (Fig. 3.43b) while the other part is in a field-free region. The sample pends on one side of a balance, the other side is loaded with a mass m. When the sample with cross section A is shifted downwards by the distance Δz due to the attractive magnetic force, the work

$$\Delta W = A \cdot M \cdot B \cdot \Delta z = \boldsymbol{F} \cdot \Delta \boldsymbol{z} = m\boldsymbol{g} \cdot \Delta \boldsymbol{z}$$

is performed against the force $\boldsymbol{F}$, where $\mathbf{M} = (\chi/\mu_0) \cdot \mathbf{B}$ is the magnetization. This work can be measured with the balance and therefore the amount of the magnetic force can be determined as

$$|\boldsymbol{F}| = (\chi/\mu_0) \cdot A \cdot B^2. \tag{3.80}$$

Example

For a sample volume $V = 1\,\mathrm{cm}^3$, a susceptibility $\chi = -10^{-6}$, a magnetic field $B = 1\,\mathrm{T}$ and a field gradient grad B = 100 T/m the force is for the Faraday method $F = 8 \times 10^{-5}$. It reaches the same value in the Gouy-method for $A = 10^{-4}\,\mathrm{m}^2$ and $B = 1\,\mathrm{T}$. One therefore needs for both methods sensitive detection techniques and strong magnetic fields.

3.5.4 Paramagnetism

The atoms of paramagnetic substances possess permanent magnetic dipole moments $\boldsymbol{p}_\mathrm{m}$. Without an external magnetic field these dipoles are randomly orientated because of their thermal motion. The average over the vector sum of the dipoles (i.e. the magnetization M)

$$\boldsymbol{M} = \frac{1}{V} \sum \boldsymbol{p}_\mathrm{m} = \mathbf{0}.$$

is therefore zero.

In an external magnetic field the dipoles are partially orientated (Fig. 3.44). The degree of orientation depends on the ratio $\boldsymbol{p}_\mathrm{m} \cdot \boldsymbol{B}/(kT)$ of potential energy of the dipole in the magnetic field to its thermal energy. In case of $\boldsymbol{p}_\mathrm{m} \cdot \boldsymbol{B} \ll k\,T$ one obtains for N dipoles per m^3 after averaging over all three spatial directions (which gives the factor 1/3) the average magnetization

$$M = N \cdot |\boldsymbol{p}_\mathrm{m}| \cdot \frac{\boldsymbol{p}_\mathrm{m} \boldsymbol{B}}{3kT} \cdot \hat{\boldsymbol{e}}_\mathrm{B} \tag{3.81a}$$

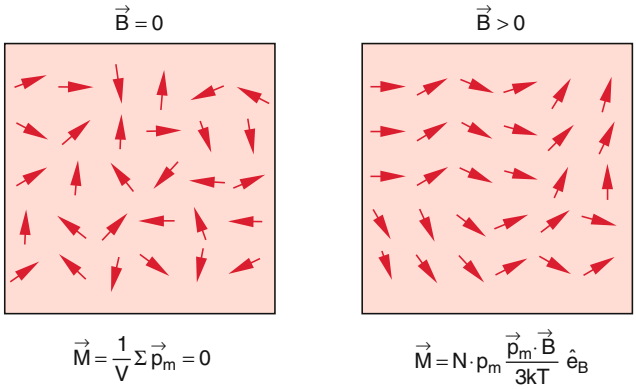


Fig. 3.44 Alignment of magnetic dipoles by an external magnetic field B which are randomly oriented at $B = 0$

With the unit vector $\hat{\boldsymbol{e}}_B$ in the field direction. This gives the susceptibility

$$\chi = \mu_0 \cdot M/B = \frac{\mu_0 \cdot N \cdot p_m^2}{3kT} \tag{3.81b}$$

This shows that χ decreases with increasing temperature as $1/T$.

3.5.5 Ferromagnetism

For ferromagnetic substances χ is very high and the magnetization can be higher by several orders of magnitude than for paramagnetic material. When a ferromagnetic sample is brought into a magnetic field B and the magnetization M is measured, one finds that $M(B)$ is not an unambiguous function of B, but depends on the previous history of the sample. If the measurement starts with a completely demagnetized sample at the external field $B = 0$, the magnetization M follows with increasing field the curve a in Fig. 3.45. Here M increases at first linearly with B and than gradually approaches saturation, where all dipoles are aligned parallel to the field $\boldsymbol{B}$.

If now the field decreases again, the magnetization $M(B)$ does not stay on the curve a but proceeds on the curve b until saturation starts again at a reverse field $-B$. After another reversal of the field the magnetization follows the curve c until it meets the curve b in the reversal point R. The

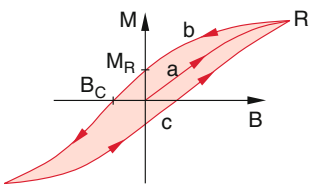


Fig. 3.45 Hysterisis loop of the magnetization $M(B)$ in an external magnetic field B

curve a is called the virginal curve, the roundtrip on b and c is the **hysteresis loop**. The residual magnetization $M_R(B = 0)$ on the curve b is the **remanence** M_R. The field strength $B = -B_c$ which is necessary to remove the residual magnetization is the **coercitivity**.

The roundtrip on the hysteresis loop requires energy for the alignment of the magnetic dipoles in the ferromagnet. In Sect. 4.4 it will be shown, that the magnetic energy in the volume V is

$$W_{magn} = \frac{1}{2} \cdot B \cdot H \cdot V. \tag{3.82}$$

The integral

$$\begin{aligned} \int M(B) \cdot dB &= \chi \cdot \mu \cdot \mu_0 \cdot \int H \cdot dH \\ &= \frac{1}{2} \cdot \chi \cdot \mu \cdot \mu_0 \cdot H^2 \\ &= \frac{1}{2} (\mu - 1) \cdot H \cdot B \end{aligned} \tag{3.83}$$

represents the area under the magnetisation curve $M(B)$ and corresponds, according to (3.82), to the magnetic energy which is required for the magnetization of the unit volume V of the sample. The area enclosed by the hysteresis loop therefore gives the energy used for one magnetization cycle. It is converted into thermal energy due to friction losses during the cycle of magnetization and demagnetization.

Most ferromagnetic substances consist of transition elements, i.e. atoms with not completely filled inner electron shells, as for example iron, nickel, cobalt, et cetera. The following experiments prove, however, that ferromagnetism is not only determined by the atomic structure but it represents a *collective phenomenon* in the solid ferromagnet, which comes about through the cooperation of many interacting atoms in the solid and would not be observed for free atoms in gases.

When a ferromagnet is heated up above a certain temperature T_c (***Curie-Temperature***) the ferromagnetism disappears. The substance remains paramagnetic for all temperatures $T > T_c$. The impressive reduction of the susceptibility χ at the Curie-Temperature can be readily demonstrated by a small iron cylinder suspended by a string which is attracted by a magnet for $T < T_c$ causing a deflection from the vertical suspension (Fig. 3.46a). When the temperature rises above T_c by heating the nail with a Bunsen burner, the nail returns to the vertical position of the suspension.

Another experiment uses a ring of ferromagnetic material, which is rotatably suspended on a pedestal with a tip (Fig. 3.46b). One part of the ring runs through a permanent horseshoe magnet. When a Bunsen burner closely behind the magnet heats the ring up above the Curie Temperature, it becomes paramagnetic. The cold ferromagnetic part of the ring is pulled into the magnetic field and the ring begins to rotate. The energy gained by pulling the ferromagnet into the

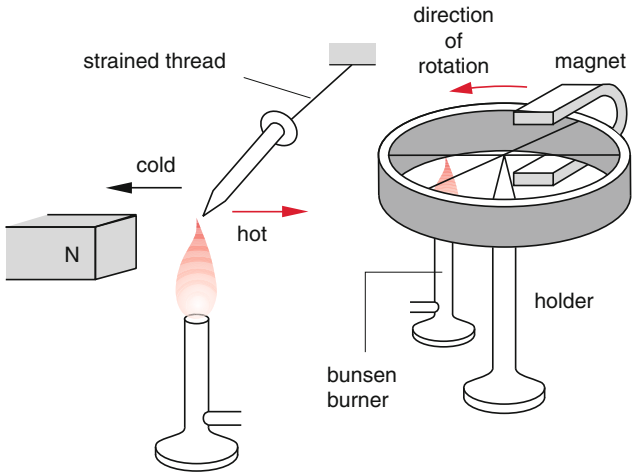


Fig. 3.46 Experimental demonstration of the disappearance of the ferromagnetism above the Curie-temperature

magnetic field is converted into kinetic energy of the rotating ring. The observed temperature dependence of the magnetic susceptibility χ can be described by the empirical formula

$$\chi(T) = \frac{C}{(T - T_C)^{\gamma}} \tag{3.84}$$

The exponent γ takes values between 1 and 1.5, depending on the material. The constant C, which also depends on the material, is the **Curie-Constant**.

In Table 3.3 the Curie-temperature T_c, the Curie constant C and the melting temperature T_{melt} are compiled. The numbers illustrate, that the Curie-Temperature (a phase transition temperature for the change from a ferro- to a paramagnet) happens already at lower temperatures than the melting temperature T_m where a phase transition from a solid into a liquid phase occurs.

When a ferromagnetic substance is evaporated, the free atoms in the gas phase are paramagnetic. This proves that a ferromagnetic solid body consists of paramagnetic atoms or molecules. The ferro-magnetism must therefore be caused by a special correlation between the atomic magnetic dipoles in the solid body.

When the magnetization curve $M(B)$ of a ferromagnetic material is measured with high resolution (i.e. the resolvable intervals ΔB are very small) it turns out that the curve $M(B)$

Table 3.3 Curie-temperature T_C, Curie-constant C and melting-temperature T_m for some ferromagnetic substances [13]

Substanz	T_C/K	C/K	T_m/K
Co	1395	2.24	1767
Fe	1033	2.22	1807
Ni	627	0.59	1727
EuO	70	4.7	1145

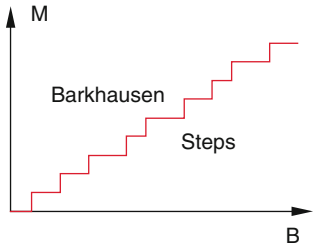


Fig. 3.47 Magnified section to the magnetization curve $M(B)$ in Fig. 3.45, which shows the steps (Barkhausen steps) which are caused by flips of magnetic areas (Weiss-areas)

does not proceed continuously, but consists of many small jumps of the magnetization (Fig. 3.47). this implies that the alignment of the atomic magnetic dipoles does not happen continuously but in small steps. These steps $\Delta M(B)$ are called *Barkhausen jumps*. They can be explained by the assumption that the ferromagnetic solid body consists of microscopic domains in which all magnetic dipoles ore orientated in the same direction due to a strong interaction between the dipoles (spontaneous magnetization). These domains which are called **Weiss domains** with the volume V_w, contain about 10^8–10^{12} atomic dipoles. Without external magnetic field the resulting magnetic moments

$$\boldsymbol{M}_w = N_w \cdot \boldsymbol{p}_m \quad N_w \approx 10^8 - 10^{12}$$

of different Weiss domains are randomly orientated. Therefore only a small total magnetic moment appears at $B = 0$ (*remanence*). When applying a magnetic field, all dipoles of a Weiss domain with volume V_w flip simultaneously into the field direction, causing a sudden increase ΔM of the magnetic moment M.

This flip occurs when the decrease of the magnetic energy

$$W_{magn} = -V_W \cdot M_W \cdot B \tag{3.85}$$

becomes larger than the energy necessary for the flip. This energy is determined by the structure of the Weiss domains and their interaction with their surroundings and can differ for the different domains. Therefore the different Weiss domains flip at different external magnetic fields.

The jumps in the magnetisation curve $M(B)$, caused by the flips of the magnetic moments in the Weiss domains can be acoustically demonstrated (Fig. 3.48). They cause sudden changes of the voltage in an induction coil surrounding the ferromagnet and can be clearly heard, when an amplifier with a loud speaker is connected to the coil.

It is possible to directly view the Weiss domains (Fig. 3.49). This can be demonstrated when a small thin iron crystal is placed in an iron-thiosulfate solution contained in a flat beaker. The sample is illuminated and the reflected light is viewed through a polarisation filter with a microscope. The polarisation of the reflected light depends on the

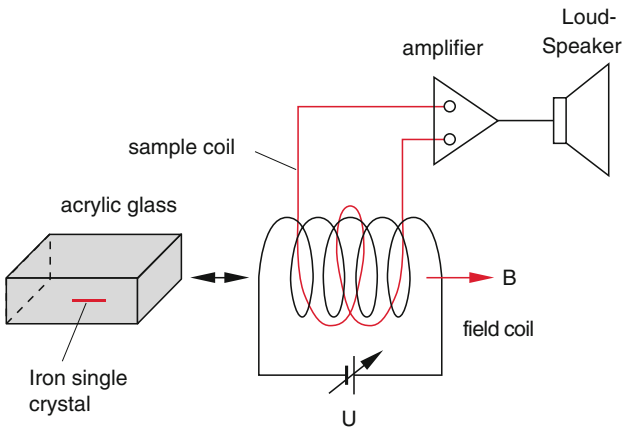


Fig. 3.48 Acoustic demonstration of the Barkhausen steps audible by short acoustic signals that are generated by a coil around the magnetic iron rod where the magnetic flips induce voltage peaks given to a loudspeaker

orientation of the magnetization in the sample. One observes bright and dark areas corresponding to the Weiss domains. When an external magnetic field is applied, some of the Weiss domains flip which appears as a brightness change of the corresponding areas (see the teaching film by Ealing about "Ferromagnetic Domain Motion" [14]).

The collective behavior of the atomic magnetic dipoles inside a Weiss domain can be formidably demonstrated by an array of small magnetic needles which are, supported by pins in a quadratic plexiglas box (Fig. 3.50). The pins are uniformly arranged in a two-dimensional quadratic or hexagonal array. When a strong magnet is moved above the assembly the magnetic needles can be randomly oriented (Fig. 3.50a). When the external magnet is removed, the small magnetic needles arrange themselves within definite areas to point all into the same direction (Fig. 3.50b), giving a macroscopic picture of the Weiss domains. With

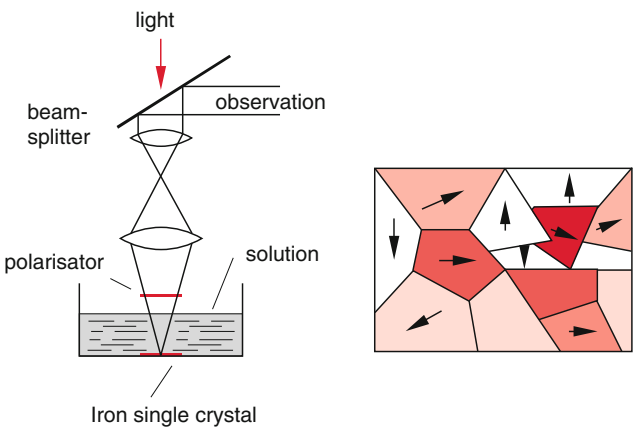


Fig. 3.49 Optical demonstration of the Weiss-areas. Polarized light is reflected by iron single crystals in solution, The plane of polarization is turned under reflection by magnetic material where the turning angle depends on the orientation of the single crystal

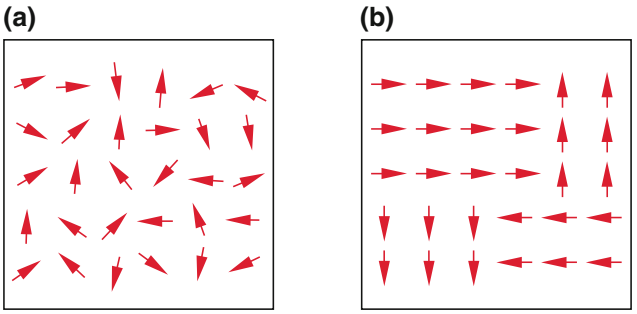


Fig. 3.50 Simple mechanical model simulating the Weiss areas by tiny magnetic needles. Without external magnetic field the needles are randomly oriented (left figure), whereas with increasing magnetic field the needles within a specific area are all oriented in the same direction, which is different for the different areas

increasing external field all magnets in more and more domains flip into the field direction. The critical field strength B_c where the flip occurs, depends on the position of the domain relative to the edge of the whole assembly and on the geometrical arrangement of the pins.

For the real ferromagnetic substances the coupling of the atomic magnetic moments, which leads to the formation of the Weiss domains is caused in a complex way by the interaction between the conduction electrons and the magnetic spin moments of atomic electrons in only partly filled inner atomic shells (see Vol. 3).

This interaction can be described by a special inneratomic magnetic field (exchange field)

$$B_e = \mu_0 \cdot \gamma \cdot M \tag{3.86}$$

which is related to the magnetization M. One can derive the relation

$$T_C = C \cdot \gamma \tag{3.87}$$

between the Curie-Temperature T_C and the Curie constant C, where γ gives the magnitude of the interaction.

Ferromagnetic substances with a strong exchange interaction show a high Curie temperature T_C.

Above the Curie-temperature T_C the thermal energy $k \cdot T$ becomes larger than the interaction energy and the ordered orientation of all magnetic moments in a Weiss domain is destroyed. The solid becomes paramagnetic.

More detailed models of ferromagnetism, which can describe all observed phenomena correctly, have been developed only recently [15, 16]

3.5.6 Antiferromagnetism, Ferri-Magnets and Ferrites

For anti-ferromagnetic substances the crystal lattice can be described by two sublattices A and B (Fig. 3.51) which have

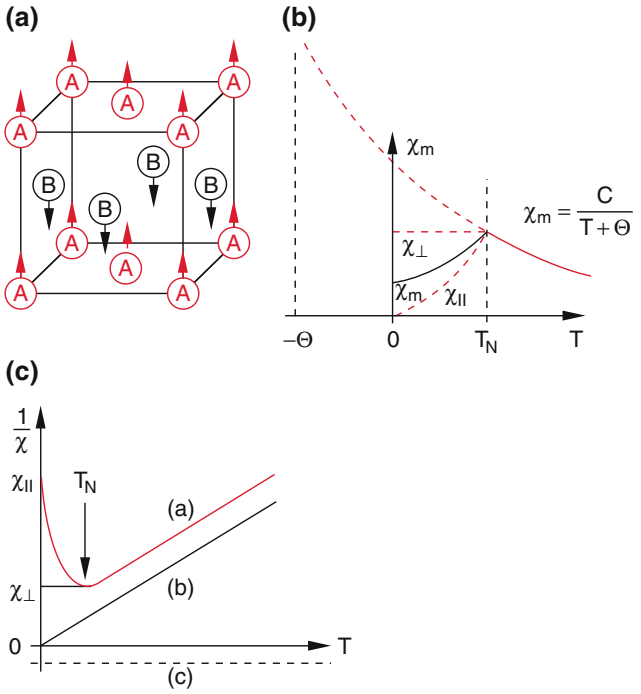


Fig. 3.51 Anti-Frerromagnet. **a**) crystal model, **b**) susceptibility, **c**) reciprocal susceptibility $1/\chi$ for antiferromagnetic material (curve **a**), paramagnetic (curve **b**) and diamagnetic (**c**) (after Dr. John Bland)

opposite directions of their atomic magnetic moments but equal amounts. Both lattices have the same number of magnetic moments and the total magnetization is therefore zero.

Examples for anti-ferromagnetic substances are metallic solids with implanted paramagnetic ions, for instance MnO, MnF_2 or UN (uranium nitride).

For ferrimagnetic material the amounts of the magnetization in the two sublattices differs. Therefore there remains a magnetization even at zero external field. By implantation of impurity atoms or molecules (e.g. Mg; Al;) into the iron lattice special ferrites can be produced which are important for many applications in electrical engineering.

The magnetization curve $M(B)$ of ferrimagnetic substances is similar to that of ferromagnets in Fig. 3.45. However, the saturation magnetization is much lower than for ferromagnets.

Quite similar to ferromagnets the ferrimagnetic materials change to paramagnets above the anti-ferromagnetic Nèel--Temperature T_N.

The susceptibility $\chi(T)$ can be described for $T > T_N$ as (Fig. 3.51b)

$$\chi = \frac{C}{T + \theta_N}. \tag{3.88}$$

where C is the Curie constant and θ_N the paramagnetic Nèel temperature.

Table 3.4 Magnetic susceptibility χ, Neel-temperature T_N and paramagnetic Neel-temperature θ_N for some antiferromagnetic substances [13]

Substance	$\chi(T_N) \cdot 10^{-9}$	T_N/K	θ_N/K
$FeCl_2$	2.5	23	+48
MnF_2	0.27	72	−113
FeO_2	0.1	195	190
MnO	0.08	120	610
CoO	0.07	291	280
Ti_2O_3	0.002	248	−2000

For anti-ferromagnets the exchange reaction can be described by the formula

$$B_{AA} = \mu_0(\gamma_{AB} - \gamma_{AA})M_A \tag{3.89a}$$

$$B_{AB} = \mu_0(\gamma_{AB} - \gamma_{AA})M_B \tag{3.89b}$$

where M_A and M_B are the magnetization of the sublattices A and B, B_{AA} is the exchange field caused by the exchange interaction between the atoms in sublattice A and B_{AB} the field due to the interaction of atomic magnetic moments in A with atoms in $\boldsymbol{B}$

For the two Nèel-temperatures T_N for antiferromagnets and θ_N for the paramagnetic Nèel temperature the relation holds

$$T_N = (C/2)(\gamma_{AB} - \gamma_{AA}) \tag{3.90a}$$

$$\theta_N = (C/2)(\gamma_{AB} - \gamma_{AA}). \tag{3.90b}$$

In Table 3.4 values of T_N and θ_N are compiled for some antiferromagnetic substances. The comparison with Table 3.3 shows that the Nèel temperatures T_N are generally distinctively lower than the Curie-temperatures T_C of ferromagnets. This demonstrates that the coupling energy, which causes the alignment of the magnetic moments is higher for ferromagnets than for antiferromagnets.

At lower temperatures $T < T_N$ antiferromagnets show collective alignments of the atomic magnetic dipoles, due to the domains structure of the sublattices. The orientation of the dipoles can be either in the field direction or perpendicular to it, depending on the orientation of the crystal structure for the different domains. This causes two different curves $\chi_{\|}(T)$ and $\chi_{\perp}(T)$, where $\chi_{\perp}$ is nearly independent of T. The geometric means $\chi_m(T) = \frac{1}{2}(\chi_{\|\|} + \chi_{\perp})$ has the temperature dependence shown in Fig. 3.51b.

In Fig. 3.52 the temperature dependent susceptibilities of paramagnetic, ferromagnetic and antiferromagnetic substances are schematically compared.

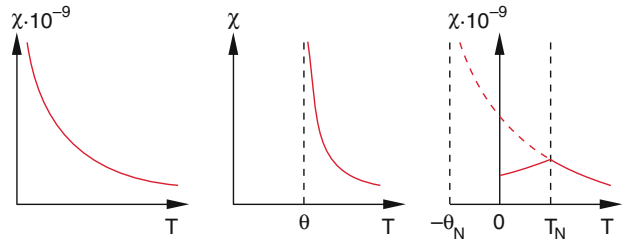


Fig. 3.52 Comparison of the temperature dependence of the susceptibilities for **a**) para- **b**) ferro- **c**) antiferro-magnetic materials

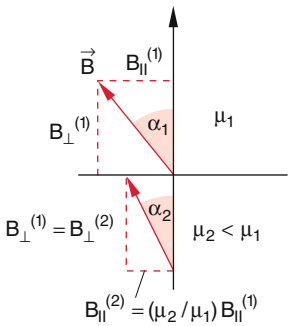


Fig. 3.53 Behaviour of normal and tangential components of the magnetic field B at the interface of two media with different permeabilities

3.5.7 Equations for the Magnetic Field in Matter

In Sect. 3.5.2 it was shown that in vacuum the relation between magnetic field strength $\boldsymbol{B}$ and magnetic intensity $\boldsymbol{H}$ is

$$\boldsymbol{B} = \mu_0 \cdot \boldsymbol{H}$$

Whereas in matter with the relative permeability constant μ the relation is

$$\begin{aligned}\boldsymbol{B} &= \mu \cdot \mu_0 \cdot \boldsymbol{H} = \mu_0 \cdot (\boldsymbol{H} + \boldsymbol{M}) \\ &= \mu_0 \cdot \boldsymbol{H} \cdot (1 + \chi)\end{aligned}$$

With the magnetization $\boldsymbol{M} = \chi \cdot \boldsymbol{H}$.

Because there are no magnetic monopoles neither in vacuum nor in matter the relation

$$\operatorname{div} \boldsymbol{B} = 0 \tag{3.91}$$

is also valid for magnetic fields in matter.

Since Ampere's law (3.5) is also valid in matter, we get for the magnetic intensity

$$\mathbf{rot}\, \boldsymbol{H} = \boldsymbol{j}, \tag{3.92}$$

where j is the density of electric currents that generate the external magnetic field $\boldsymbol{B}_a = \mu_0 \cdot \boldsymbol{H}$.

For the field B in homogeneous matter we obtain from div $\boldsymbol{B}$ = 0

$$\begin{aligned}\operatorname{div} \boldsymbol{B} &= \operatorname{div}(\mu \cdot \mu_0 \cdot \boldsymbol{H}) \\ &= \mu \cdot \mu_0 \cdot \operatorname{div} H + \mu_0 \cdot H \cdot \operatorname{grad} \mu = 0.\end{aligned}$$

In homogeneous substances is grad μ = 0 and therefore div $\boldsymbol{H}$ = 0, while in inhomogeneous media grad $\mu \neq 0$ and therefore generally div $\boldsymbol{H} \neq 0$.

In Sect. 1.7.3 the behaviour of the vectors $\boldsymbol{E}$ and $\boldsymbol{D}$ at the boundary of two media with different permittivities ε have been discussed. It turned out, that for the transition from medium 1 into medium 2 the tangential component of $\boldsymbol{E}$ is continuous ($E_{\parallel}^{(1)} = E_{\parallel}^{(2)}$) whereas the perpendicular component is discontinuous ($E_{\perp}^{(1)} = (\varepsilon_2/\varepsilon_1) \cdot E_{\perp}^{(2)}$). The behaviour of $\boldsymbol{D}$ is just opposite.

A similar behaviour is found for the magnetic vectors $\boldsymbol{B}$ and $\boldsymbol{H}$. The argumentation is completely analogous to that in Sect. 1.7. One can conclude that in media without electric currents ($\boldsymbol{j}$ = 0) the condition **rot** $\boldsymbol{H}$ = **0** holds. This implies (analogous to **rot** $\boldsymbol{E}$ = **0**) that the tangential component of $\boldsymbol{H}$ is continuous at the boundary between a medium with $\mu = \mu_1$ and one with $\mu = \mu_2$.

$$H_{\parallel}^{(1)} = H_{\parallel}^{(2)} \Rightarrow \frac{B_{\parallel}^{(1)}}{\mu_1} = \frac{B_{\parallel}^{(2)}}{\mu_2}. \tag{3.93a}$$

For the perpendicular component we get from div B = 0 (see Problem 3.10) the condition

$$B_{\perp}^{(1)} = B_{\perp}^{(2)} \Rightarrow \mu_1 H_{\perp}^{(1)} = \mu_2 H_{\perp}^{(2)}. \tag{3.93b}$$

Similar to Snellius law of refraction in optics we can derive from (3.93a, 3.93b) a refraction law for magnetic fields which describes the direction change of the magnetic field vector $\boldsymbol{B}$ at the boundary between two media with different values of μ (Fig. 3.53):

$$\tan \alpha_1 = B_{\parallel}^{(1)}/B_{\perp}^{(1)} \text{ and } \tan \alpha_2 = B_{\parallel}^{(2)}/B_{\perp}^{(2)}$$

which gives for the change of the angle α

$$\frac{\tan \alpha_1}{\tan \alpha_2} = \frac{\mu_1}{\mu_2}. \tag{3.94}$$

3.5.8 Electromagnets

The enhancement of the magnetic field B by substances with a high value of the permeability μ is used in electromagnets. Their principle can be explained as follows:

The inside of a toroidal solenoid with N windings carrying a current I is wrapped around an iron ring (Fig. 3.54a). For a closed integration loop in the iron ring we get, according to (3.6) and (3.93a) the condition

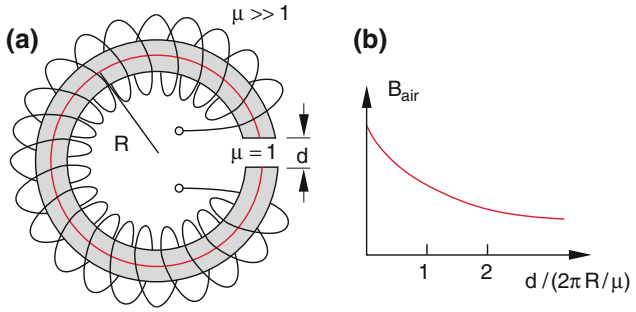


Fig. 3.54 **a**) Circular coil with iron core and air gap of thickness d. **b**) Magnetic field B as a function of d

$$\int H \cdot \mathrm{d}s = 2\pi \cdot R \cdot H = N \cdot I.$$

This gives

$$H = \frac{N \cdot I}{2\pi \cdot R} \Rightarrow B = \mu \cdot \mu_0 \cdot \frac{N \cdot I}{2\pi \cdot R}. \tag{3.95}$$

Now we consider an iron ring with a small air gap with width d. Since the normal component of B is continuos at the boundary iron-air the condition holds with $\mu_{\mathrm{air}} = 1$

$$B_{\mathrm{iron}} = B_{\mathrm{air}} \Rightarrow \mu \cdot H_{\mathrm{iron}} = H_{\mathrm{air}} \tag{3.96}$$

For the line integral over the magnetic intensity H we obtain for one circulation

$$\begin{aligned} N \cdot I &= \int \boldsymbol{H} \cdot \mathrm{d}s = (2\pi \cdot R - d) \cdot H_{\mathrm{iron}} + d \cdot H_{\mathrm{air}} \\ &= \left(\frac{2\pi \cdot R - d}{\mu} + d \right) \cdot H_{\mathrm{air}}. \end{aligned} \tag{3.97}$$

With (3.6) the magnetic intensity in the air gap becomes

$$\begin{aligned} H_{\mathrm{air}} &= \frac{N \cdot I \cdot \mu}{(\mu - 1)d + 2\pi R} \\ &\approx \frac{N \cdot I \cdot \mu}{\mu \cdot d + 2\pi R} \quad \text{for } \mu \gg 1 \end{aligned} \tag{3.98}$$

And for the magnetic field strength

$$B_{\mathrm{air}} = \frac{\mu \cdot \mu_0 \cdot N \cdot I}{\mu \cdot d + 2\pi R}. \tag{3.99}$$

For a gap width $d = 2\pi R/\mu$ the field strength B has decreased to one half of its value in the iron core. Since the permeability of iron is about $\mu = 2000$, the values of B and H in the air gap decrease rapidly with increasing width d of the air gap (Fig. 3.54b)

Example

With a toroidal solenoid with iron core ($\mu = 2000$), a radius $R = 20$ cm and $N = 5000$ windings carrying a current of $I = 1$ A, a magnetic field $B = 0.6$ T can be generated in an air gap with $d = 1$ cm. Increasing the gap to $d = 2$ cm decreases the maximum field already to 70% or 0.42 T.

3.6 The Magnetic Field of the Earth

The magnetic field of the earth has been used for navigation with compass needles for more than 2000 years. The insight that needles of the mineral magnetite (the name comes from the city magnesia in Turkey) always points to the north was known to the Greek for about 1500 years and even earlier for the Chinese. The exact form of the earth's magnetic field was, however, measured not before the 19th century and definite models about its origin and its variation in time have been developed only in the 20th century. Even today many details are still unclear.

The magnetic field of the earth can be approximately described by the field of a magnetic dipole in the center of the earth. The dipole axis is at present inclined by 11.4° against the rotation axis of the earth (Fig. 3.55). The amount of the dipole moment is $1/\gamma^2$ [17]. The total field strength on the surface of the earth is 25 μT at the equator and 70 μT at the poles. The horizontal and vertical components of the magnetic field on the earth surface also depend on the geographic latitude φ. The total energy of the magnetic field outside the earth is 10^{18} J, that inside the earth is about two orders of magnitude higher.

People have named that pole of the magnetic needle that points to the north as north-pole.

Since a magnetic north pole is attracted by a magnetic south pole, the earth magnetic pole close to the geographic north-pole should be named magnetic south-pole. In order to avoid confusion, nowadays this pole is named arctic *pole* while the magnetic north-pole at the geographic south-pole is named *antarctic pole*. The two magnetic poles do not coincide with the geographic poles (the points where the rotation axis pierces through the surface of the earth), but show a small deviation, which changes slowly with time.

More detailed measurements have shown, that the actual magnetic field of the earth deviates slightly from a dipole field B_{D}. The difference between the real field and the dipole field

$$\Delta B(\theta, \varphi) = B_{\mathrm{r}}(\theta, \varphi) - B_{\mathrm{D}}(\theta, \varphi)$$

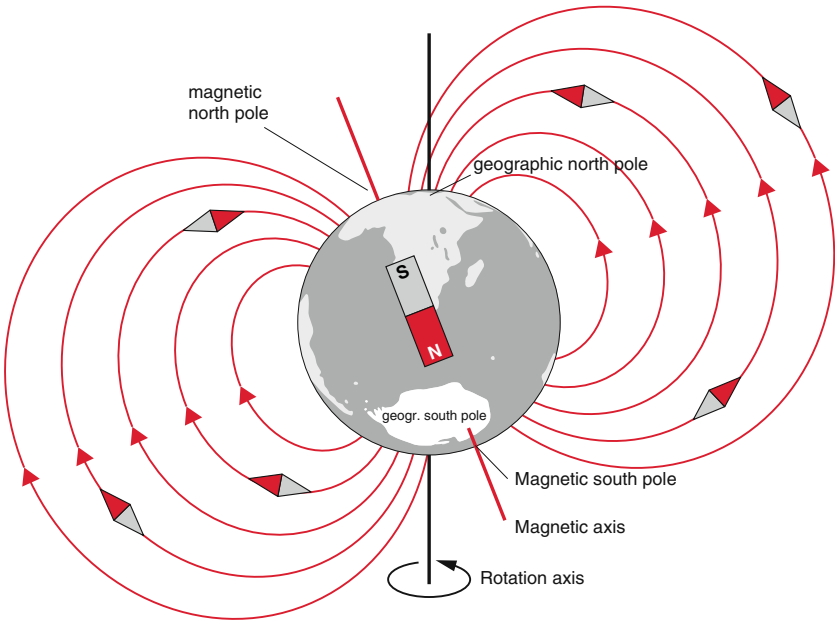


Fig. 3.55 Magnetic field of the earth

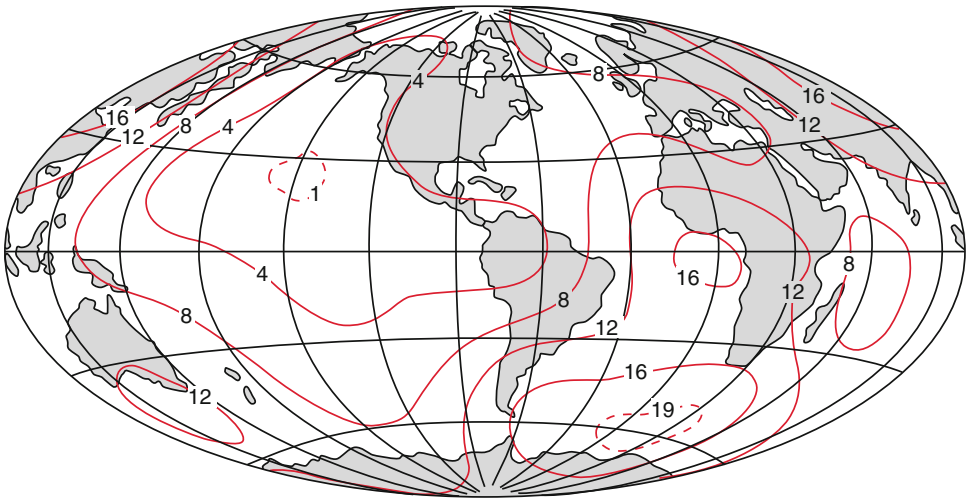


Fig. 3.56 Deviation of the measured magnetic field from a pure dipole field. The curves connect locations with the same deviation given in units of micro-Tesla

on the earth surface depends on the geographical longitude θ and latitude φ. In Fig. 3.56 the curves of equal ΔB values given in μT are shown. The local variations are caused by different effects. One of them is the nonuniform distribution of magnetic minerals in the earth crust. Whereas the field strength of the dipole field decreases for $r > R$ with $1/r^3$ the difference ΔB declines with $1/r^4$. Therefore the magnetic field of the earth approximates with increasing r more and more an ideal dipole field.

Far away from the earth in the interplanetary space the dipole field is strongly altered by currents of charged particles (electrons and proton) emitted from the sun (solar wind, see Vol. 4) [18] and Fig. 3.57. The magnetic field shields the earth surface (and therefore mankind) from the solar wind, which can only enter along the magnetic lines immersing into the earth close to the poles, where they cause the *aurora borealis* (Northern Light) by collisions with molecules in the air.

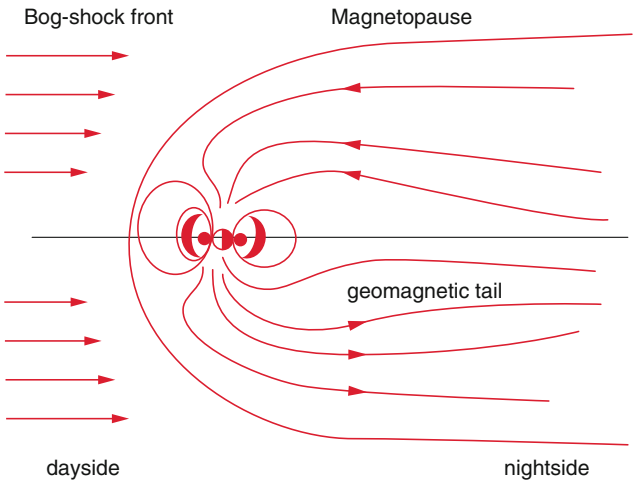


Fig. 3.57 Deformation of the earth' s magnetic field by the sun wind

The magnetic poles wander slowly in the course of time (Fig. 3.58). One distinguishes between two different poles: The magnetic pole (this is the location where the magnetic field lines pierce vertically through the earth surface) and the geo-magnetic pole (this is the point where the axis of the magnetic dipole moment touches the earth surface).

An important experimental result is the decrease of the earth's magnetic field with time. It turned out that magnitude and direction of the field change in course of time (Fig. 3.59). Investigations of the magnetization of ferromagnetic minerals from volcanoes and in the sediments of the ocean ground, where permanently magma is supplied from the interior of the earth, conclusions about the variation of the magnetic field during geological periods of time can be obtained. Such conclusions are based on the assumption that in the magnetic minerals the orientation of the magnetization following during their liquid phase the earth magnetic field and that this orientation was fixed when the minerals solidified and has been no more altered. The time of the solidification can be determined by measurements of the sequence of geological layers and by radioactive dating methods (see Vol. 4).

It turns out, that the earth's magnetic field reverses its direction in random time intervals. The average time between successive field reversions for a specific field reversion is about 2×10^5 years. The overturn time itself is much shorter. The magnetic field breaks down in about 10^4 years and rebuilds itself in the reverse direction in about the same time.

The question is now, what is the origin of the magnetic field, which mechanism generates this field?

Since all ferromagnetic minerals in the interior of the earth have a Curie-temperature T_C which is below the temperature in the earth's interior, these minerals cannot be the source of the magnetic field.

Note For temperatures above the Curie temperature T_c ferromagnets change their ferromagnetism into the much weaker paramagnetism (see Sect. 3.5.5 and Books on Solid State Physics).

Also the random fluctuations of the field with time exclude permanent magnets, i.e. magnetic minerals as the source of the field. These magnetic minerals in the earth crust only cause small local variations of the magnetic field. Therefore the assumption is that the magnetic field is generated by electric current loops in the interior of the earth around the axis of the magnetic dipole where the high temperature causes the liquidification of all materials (magma). These liquids contain ionized atoms and they move inside the liquid region of the earth's interior.

The question is now: "what drives these liquids to flow? There are several possible causes:

- Due to the radial temperature gradient convection currents are generated. Liquid material rises upwards, cools down and sinks down again, because the cold material has a higher density. The rotation of the earth causes a Coriolis force $F_C = 2m \cdot (v \times \omega)$ on a particle with mass m moving in the radial direction and deflects the convection current into the tangential direction.
- Because of the missing restoring forces of liquids the liquid region of the rotating earth shows a larger centrifugal widening at the equator and a larger pole oblateness as the solid crust. Therefore the principal inertial axis of the liquid region does not revolute on the same precession cone as the earth rotation axis. The torques causing the precession (see Vol. 1, Chap. 5) are different for the liquid and the solid region. This leads to a relative motion of the liquid against the solid region and causes magma currents. For such currents of partly ionized materials the total electric current density

$$\boldsymbol{j} = \varrho^+ \boldsymbol{v}^+ + \varrho^- \boldsymbol{v}^-$$

depends on the different flow velocities of positively and negatively charged particles, since the negative charge carriers are generally electrons which have a much higher mobility and therefore a higher drift velocity. The magnetic field, generated by these electric currents causes a Lorentz force $\boldsymbol{F}_L = q \cdot (\boldsymbol{v} \times \boldsymbol{B})$ which separates positive and negative charges and increases the difference of the drift velocities. This leads to an amplification of the magnetic field. The motion of the charge carriers generates an additional magnetic field which reinforces the initial field. This is illustrated in Fig. 3.60 by an experimental example.

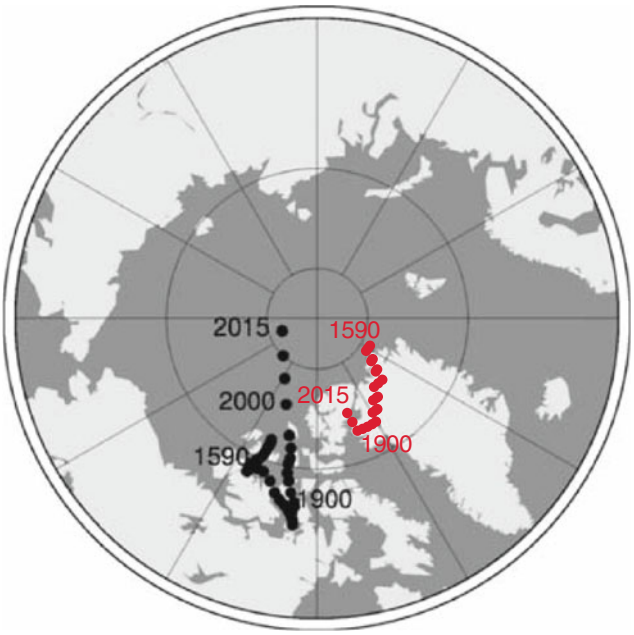


Fig. 3.58 Migration of the northern magnetic pole (black) and the geo- magnetic pole (red)

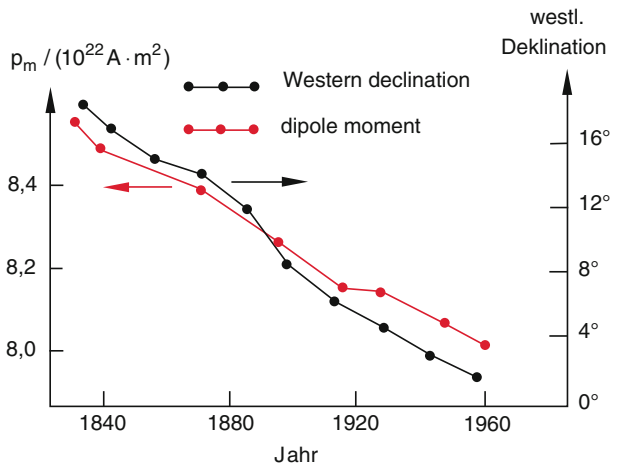


Fig. 3.59 Temporal change of magnitude and direction of the earth's magnetic field in Frankfort, Germany

An electrically conducting disc rotates about the axis A in a magnetic field with $\boldsymbol{B}||\boldsymbol{A}$. When two electric sliding contacts S_1 and S_2 are connect by a conductor loop an electric current flows through the loop which generates a magnetic field that is parallel to the original field and therefore reinforces it (***dynamo principle***) (Fig. 3.60).

Due to friction losses and emerging turbulence in the motion of the magma the currents can change in time. They can even have a spatial distribution of the currents which produces no magnetic field and they can change their directions.

Many details of this model are still not clear and more investigations are necessary before a complete understanding of the origin of the earth's magnetic field is reached [19].

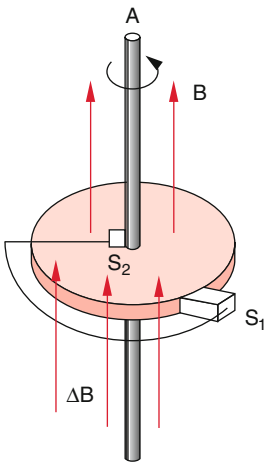


Fig. 3.60 Dynamo-model for the explanation for the amplification of the magnetic field by electric currents, driven by the Lorentz force

Summary

- Magnetic fields can be generated by permanent magnets and by electric currents. Between the magnetic field strength $\boldsymbol{B}$ and the magnetic intensity $\boldsymbol{H}$ exists the relation $\boldsymbol{B} = \mu_0 \cdot \boldsymbol{H}$ (μ_0 = permeability constant in vacuum.)
- Stationary magnetic fields are source-free (div $\boldsymbol{B}$ = 0). This implies that there are no magnetic monopoles.
- For a closed path around a conductor carrying the electric current $I = \int \boldsymbol{j} \cdot \mathrm{d}\boldsymbol{A}$ the relation holds

$$\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{s} = \mu_0 \cdot I \Rightarrow \mathbf{rot}\,\boldsymbol{B} = \mu_0 \cdot \boldsymbol{j}.$$

- The magnetic field around a straight wire with radius R conducting the current I is cylindrical symmetric and the radial dependence is

$$B(r) = \frac{\mu_0 \cdot I}{(2\pi r)} \quad \text{for } r > R \quad \text{and} \quad B(r) = \frac{\mu_0 \cdot j \cdot r}{2} \quad \text{for } r < R$$

- The magnetic field in the interior of a long solenoid with n windings per m and the current I is homogeneous and has the amount

$$B = \mu_0 \cdot n \cdot I$$

- The vector potential $\boldsymbol{A}$ of a magnetic field B is related to B by

$$\boldsymbol{B} = \mathbf{rot}\,\boldsymbol{A}.$$

  $\boldsymbol{A}$ can be defined unambiguously by the additional condition (Coulomb gauge)

$$\mathrm{div}\,\boldsymbol{A} = 0.$$

- The vector potential $\boldsymbol{A}(\boldsymbol{r_1})$ at the point $\boldsymbol{r_1}$ outside an arbitrary current distribution $\boldsymbol{j}(\boldsymbol{r_2})$ within the volume V_2 is

$$\boldsymbol{A}(\boldsymbol{r_1}) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{j}(\boldsymbol{r_2}) \cdot (\mathrm{d}V_2)}{|\boldsymbol{r_1} - \boldsymbol{r_2}|}.$$

- The Lorentz force acting on a charge q moving with the velocity $\boldsymbol{v}$ in a magnetic field $\boldsymbol{B}$ superimposed by an electric field $\boldsymbol{E}$ is

$$\boldsymbol{F} = q(\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \quad (Biot - SavartLaw)$$

- The force acting on the length $\mathrm{d}L$ of a conductor with current I in a magnetic field $\boldsymbol{B}$ is

$$\boldsymbol{F} = I(\mathrm{d}\boldsymbol{L} \times \boldsymbol{B}).$$

- Axial magnetic fields act as lenses for focusing a beam of charged particles.
- Homogeneous magnetic sector fields can be used to separate charged particles with different masses. They form the basis of magnetic mass spectrometers.
- Metals and semiconductors with current density j and electron charge density $n \cdot q$ generate in a magnetic field $\boldsymbol{B}$ the Hall voltage

$$U_H = -(\mathbf{j} \times \mathbf{B}) \cdot \mathbf{b}/(n \cdot q)$$

  where $\boldsymbol{b}$ is the thickness of the probe perpendicular to $\boldsymbol{B}$.
- The magnetic field $\boldsymbol{B}$ of an electric current and the Lorentz force acting on a charge q moving in a magnetic field can be deduced by the relativity theory just from the Coulomb law and the Lorentz transformations.
- The magnetic part $q \cdot (\boldsymbol{v} \times \boldsymbol{B})$ of the Lorentz force can be attributed to electric forces by transformation to a moving inertial system. In this system the magnetic field becomes zero.
- The electric field as well as the magnetic field generally alter in different inertial systems However, the total force and with it the equations of motion remain invariant.
- The magnetic properties of materials are described by the magnetic susceptibility χ. One distinguishes between

$$\begin{array}{ll} \text{Diamagnets :} & |\chi| \ll 1; \quad \chi < 0 \\ \text{Paramagnets :} & |\chi| \ll 1, \quad \chi > 0 \\ \text{Ferromagnets :} & |\chi| \gg 1; \quad \chi > 0 \\ \text{Antiferromagnets :} & 0 < |\chi| < 100; \end{array}$$

  For paramagnets the temperature dependence of the susceptibility χ can be described by $\chi = \mathrm{C}/(\mathrm{T} + \theta)$ where θ is the Neel temperature. The susceptibility χ is smaller than for ferromagnets.
- In matter the relation between $\boldsymbol{B}$ and $\boldsymbol{H}$ is

$$\boldsymbol{B} = \mu \cdot \mu_0 \cdot \boldsymbol{H} = \mu_0 \cdot (1 + \chi) \cdot \boldsymbol{H}.$$

  The dimensionless constant μ is the relative permeability number.
- The magnetic dipole moment of the area $\boldsymbol{A}$ enclosed by the current I is defined as $\boldsymbol{p}_\mathrm{m} = I \cdot \boldsymbol{A}$
- The magnetization

$$\boldsymbol{M} = \chi \cdot \boldsymbol{H} = \frac{1}{V} \sum \boldsymbol{p}_\mathrm{m}$$

  gives the vector sum of all atomic magnetic dipoles per volume V.
- Ferromagnetism is determined by the macroscopic structure of specific magnetic matter and depends on the order of the atomic dipole orientations. It disappears above the Curie-temperature.
- The magnetic field of the earth is mainly caused by electric currents of the liquid magma in the interior of the earth. Magnetic minerals in the earth crust cause only small local variations of the magnetic field.

Problems

3.1 Two long straight wires are stretched in the z-direction with a mutual distance of 2a = 2 cm. Each wire carries a current of 10 A in the same direction and also in the opposite direction.
- (a) Illustrate the magnetic field in the x-y-plane by drawing the field lines.
- (b) Calculate the magnetic field on the x- as well as on the y-axis (see Fig. 3.61a).
- (c) Calculate the forces per m length between the two wires.
- (d) What is the force, if the two wires are perpendicular to each other, i.e. one wire lies in the line $z = y = 0$, the other in the line $x = 0$, $y = -2$ cm (Fig. 3.61b)?

3.2 Two concentric tubes carry a constant current I flowing into opposite directions through the two tubes (Fig. 3.62). Determine the magnetic field and its dependence on the distance r from the axis ($0 \leq r \leq \infty$).

3.3 In the hydrogen atom the electron ($m = 9.109 \times 10^{-31}$ kg, $e = 1.6.2 \times 10^{-19}$ C) moves according to the Bohr model on a circle with radius $r = 0.529 \times 10^{-10}$ m around the nucleus. Which average electric current corresponds to this electron motion and what is the magnetic field at the nucleus?

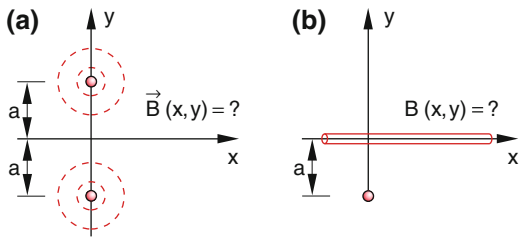


Fig. 3.61 Illustration of Problem 3.1

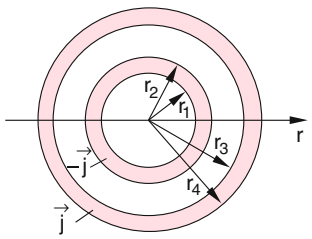


Fig. 3.62 Illustration of Problem 3.2

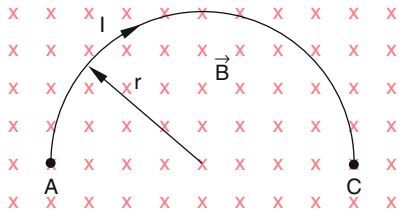


Fig. 3.63 Illustration of problem 3.4

3.4 In a plane perpendicular to a magnetic field lies a wire on a half circle (Fig. 3.63) which carries the current I. Show that the force on the wire is the same as that acting on a straight wire along the line AC between the ends of the half circle.

3.5 Consider a Helmholtz coil pair with a radius of 40 cm and 100 windings of each coil. The current I flows in both coils into the same direction.
- (a) What is the magnetic field in the center at $z = 0$ when the coil distance is d?
- (b) What is the current I in order to compensate the earth magnetic field of 5×10^{-5} T = 0.5 Gauß? What is the orientation of the coil axis for a complete compensation?
- (c) What is the dependence $B(z)$ on the coil axis outside of the coils?

3.6 An electron starts from $P = \{0, 0, 0\}$ with the velocity $(v_0/\sqrt{3})\{1, 1, 1\}$ in a homogeneous magnetic field $\boldsymbol{B} = \boldsymbol{B}_0 \cdot \{0, 0, 1\}$
- (a) Describe the path of the electron
- (b) How does the path change, when an electric field $\boldsymbol{E_1} = E_0 \cdot \{0, 0, 1\}$ or $\boldsymbol{E_2} = E_0 \cdot \{1, 0, 0\}$ is superimposed?
- (c) Which of the following quantities are not affected by the electric field E_1 or E_2: $v_x, v_y, v_z, v_r, |\boldsymbol{v}|\,|\boldsymbol{p}|, \boldsymbol{p}, E_{\text{kin}}$?

3.7 A thin copper rod with rectangular cross section (thickness $\Delta x = 0.1$ mm width $\Delta y = 1$ cm) carrying a current $I = 10$ A is oriented perpendicular to a magnetic field $\boldsymbol{B} = \{B_x, 0, 0\}$ with 2 T. Assuming that each copper atom delivers a free electron ($n_e = 8 \times 10^{22}/\text{cm}^3$) calculate
- (a) The drift velocity of the electrons
- (b) The Hall Voltage
- (c) The force per m on the copper rod.

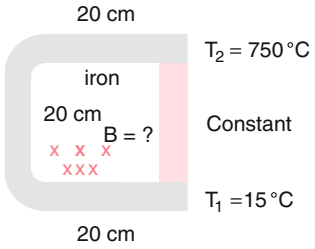


Fig. 3.64 Illustration of problem 3.8

3.8 A rod of constantan (length L = 20 cm, cross section A = 5 mm^2 specific resistance $\rho = 0.5 \times 10^{-6}$ ohm $\cdot\, m$ and an iron yoke (L = 60 cm, A = 5 mm^2, $v = \{v_x, 0, 0\}$) are soldered together on both sides of the constantan rod (Fig. 3.64). The constant a of the thermo-voltage in Eq. (2.42a) is a = 53 μ V/K.
   (a) Calculate the current I if one end is in water at T = 15 °C and the other end is heated by a flame to T = 750 K.
   (b) Calculate the magnetic field in the center of the quadratic loop.

3.9 Calculate the velocity interval Δv which is transmitted when a parallel beam of charged particles with velocity v_0 passes through a Wien filter with a slit width Δb (Fig. 3.27).

3.10 Show, that Eq. (3.93b) follows from the condition div $\boldsymbol{B}$ = 0.

References

1. M. Spiegel: Vector Analysis 2nd ed. (McGraw Hill 2009)
2. R.L. Schilling: measures, Integrals and Martingales (Cambridge Univ. Press 2017)
3. H. Hancock: Elliptical Integrals (Dover Publcations 1958)
4. P.F. Byrd, M.D. Friedmann: Handbook on Elliptical Integrals for Engineers and Physicists (Springer Heidelberg 1954)
5. https://en.wikipedia.org/wiki/Magneto-optical_trap
6. Th.W. Burgoyne, Gary M. Hieftje (1996). "An introduction to ion optics for the mass spectrograph". Mass Spectrometry Reviews. **15** (4): 241–259
7. https://en.wikipedia.org/wiki/Mass_spectrometry
8. J.H. Gross: Maas Spectrometry A textbook (3rd ed. Springer 2017)
9. https://en.wikipedia.org/wiki/Hall_effect_sensor#Hall_probe
10. John Cogut; Special Relativity, Electrodynamics and General Relativity (Academic Press 2018)
11. http://www.damtp.cam.ac.uk//user/db275/concepts/EM.pdf
12. CRC Handbook of Chemistry and Physics 99th ed. (chemical Rubber company press, Taylor and Francis 2018)
13. H. Stöcker: Taxchenbuch der Physik (Verlag Harri Deutsch Frankfurt 1994)
14. Ealing Single Concept Films. http://www.phy.mtu.edu/LECDEMO/websit/ealing.html
15. https://en.wikipedia.org/wiki/Ferromagnetism
16. P.J. Wyn & Henricus: Ferromagnetism (Encyklopedia of Physics, XIII, 2 (Springer Heidelberg) Amikam Maharoni: Ferromagnetism (Oxford Univ. Press 2001)
17. https://en.wikipedia.org/wiki/Earth%27s_magnetic_field. http://hyperphysics.phy-astr.gsu.edu/hbase/magnetic/MagEarth.htm. http://www.revimage.org/earth-magnetic-field-strength-map
18. https://en.wikipedia.org/wiki/Solar_wind
19. DE. Dormy: The origin of the earth magnetic field. Europhysi9cs News **37**, 22 (2016)

Temporally Variable Fields

4

Up to now we have treated only temporally constant electric and magnetic fields. All properties of these static fields caused by resting charges or stationary currents can be derived from a few basic equations (see Chaps. 1–3). These equations are based on experimental observations and are:

$$\begin{aligned} \mathbf{rot}\,\boldsymbol{E} &= 0 \qquad & \mathbf{rot}\,\boldsymbol{B} &= \mu_0 \cdot j \\ \operatorname{div}\boldsymbol{E} &= \varrho/\varepsilon_0 & \operatorname{div}\boldsymbol{B} &= 0 \\ \boldsymbol{E} &= -\mathbf{grad}\,\phi & \boldsymbol{B} &= \mathbf{rot}\,\boldsymbol{A}. \\ & \boldsymbol{j} = \sigma \cdot \boldsymbol{E} \end{aligned} \tag{4.1}$$

From the spatial distribution of charges $\varrho(x, y, z)$ the electric field strength $\boldsymbol{E}(x, y, z)$ and the electric potential $\phi(x, y, z)$ can be calculated while from the current distribution $\boldsymbol{j}(x, y, z)$ the magnetic field strength $\boldsymbol{B}$ and the vector potential $\boldsymbol{A}$ can be obtained. The connection between $\boldsymbol{j}$ and $\boldsymbol{E}$ is given by the electrical conductivity σ as a material constant of the respective conductor. As has been shown in (3.60) the constants of nature ε_0 and μ_0 are connected to the speed of light c in vacuum by

$$\varepsilon_0 \cdot \mu_0 = 1/c^2$$

Now the question is how these equations have to be extended, if charge density ϱ and current density $\boldsymbol{j}$ and with it also electric and magnetic fields are temporally changing.

In this chapter we consider "slow" temporal changes where the time Δt, which light needs to travel over the area of the distribution of charges or currents is very short compared to the time interval Δt of the temporal change of ϱ resp. $\boldsymbol{j}$ so that we can neglect the changes within Δt. In Chap. 6 we drop this restriction.

4.1 Faraday's Law of Induction

Michael Faraday (Fig. 4.1) was the first who recognized that an electric voltage is generated across a conductor in a temporally variable magnetic field. He named this voltage *induction voltage*. At first we will discuss some fundamental

Fig. 4.1 Michael Faraday (1791–1867) found 1831 the induction law named after him

experiments to show the quantitative connection between this induction voltage and the temporal change of the magnetic flux.

1. The north pole of a bar magnet is pushed through a coil with N windings with its endings connected to an oscillograph to measure the temporally varying voltage (Fig. 4.2). During the motion of the magnet one observes a voltage $U(t)$. Its amount and temporal course depends on several factors. $U(t)$ is proportional
   (a) to the speed $v(t)$ with which the magnet is moved through the coil,
   (b) to the product $N \cdot A$, of the number of windings of the coil and its cross sectional area A,
   (c) to the cosine of the angle a between the surface normal $\boldsymbol{A}_N$ of the coil and the magnetic field direction $\boldsymbol{B}$.

   If the experiment is performed with the south pole of the magnet, the observations are the same but the voltage has the opposite polarity.
2. In the homogeneous field of a Helmholtz-coil the cross section A of a flat flexible test coil with N windings is reduced by ΔA when compressing the coil. Again one

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_4

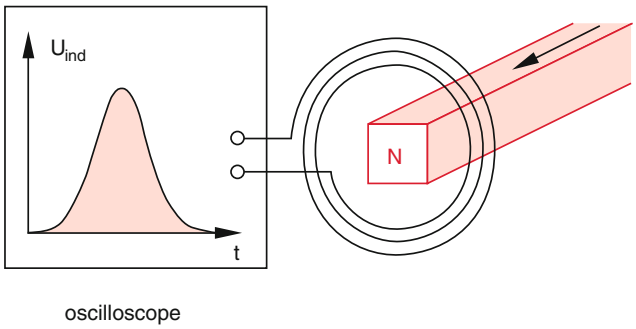


Fig. 4.2 When a magnetic rod is pushed through a solenoid with N windings an electric voltage is $U_{ind}(t)$ is measured between the ends of the solenoid, which is proportional to the time derivative $d\Phi_m/dt$ of the magnetic flux through the solenoid

observes an induction voltage whose amount depends on the speed of the surface change $\Delta A(t)$.

3. Instead of the bar magnet in the first experiment a current carrying cylindrical solenoid is used that has n windings per unit length (see Sect. 3.2.6). Its magnetic field can be changed by changing the current. A small test coil, revolving about a vertical axis inside the field coil, detects the induction voltage.

The induction voltage $U(t)$ as well as the current $I(t)$ of the field coil and with it the magnetic field $B(t) = \mu_0 \cdot n \cdot I(t)$ are displayed on a double beam oscilloscope (Fig. 4.3). Now we run a current $I(t) = I_0 \cdot \sin \omega t$ through the field coil with total area $N \cdot A$. The test coil remains at rest but we vary the frequency ω and get the induction voltage

$$U_{ind} = U_0 \cdot \sin(\omega t + 90^\circ)$$

With

$$U_0 = -\omega \cdot B \cdot N \cdot A \cdot \cos \alpha,$$

N is the number of windings of the test coil, A is its cross sectional area, and α the angle between the normal vector $\boldsymbol{A}_N$ and the direction of the field $\boldsymbol{B}$.

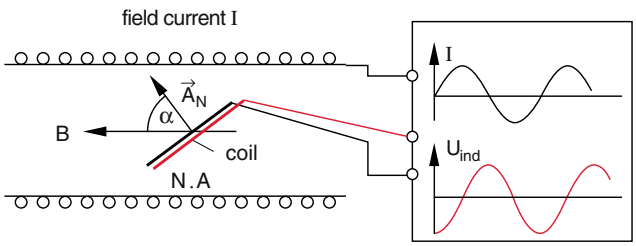


Fig. 4.3 Induction voltage between the ends of a fixed coil with N windings and area A in a temporal changing magnetic field

The results of these three experiments show that the measured induction voltage is the negative temporal change of the magnetic flux through the test coil. This is Faraday's law:

$$U_{ind} = -\frac{d}{dt}\int \boldsymbol{B} \cdot d\boldsymbol{A} = -\frac{d\Phi_m}{dt}. \qquad (4.2)$$

Examples

1. A rectangular coil with N windings and cross section A rotates with constant angular velocity ω in a constant homogeneous magnetic field $\boldsymbol{B}_0$ (Fig. 4.4). Then the magnetic flux Φ_m through the coil is

$$\Phi_m = \int \boldsymbol{B} \cdot d\boldsymbol{A} = B \cdot N \cdot A \cdot \cos \varphi(t) \qquad (4.2a)$$

where $\varphi(t) = \omega \cdot t$ is the angle between the normal vector and the direction of the field. According to (4.2) the induced voltage is then

$$U_{ind} = -\frac{d}{dt}\Phi_m = B \cdot N \cdot A \cdot \omega \cdot \sin \omega t. \qquad (4.2b)$$

This equation represents the fundamentals of the technical realization of alternating current generators. Its basic principle can be demonstrated by a simple model of an ac-generator driven by hand (Fig. 4.5). A few realistic technical realizations are explained in Chap. 5.

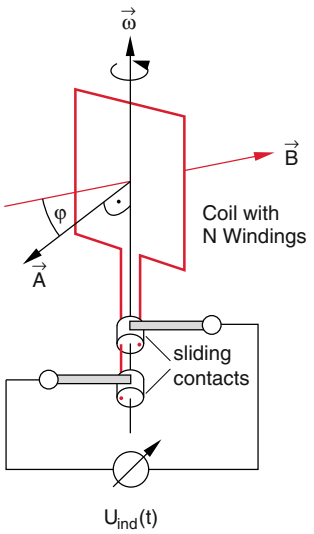


Fig. 4.4 Generation of an ac induction voltage by turning a coil in a constant magnetic field

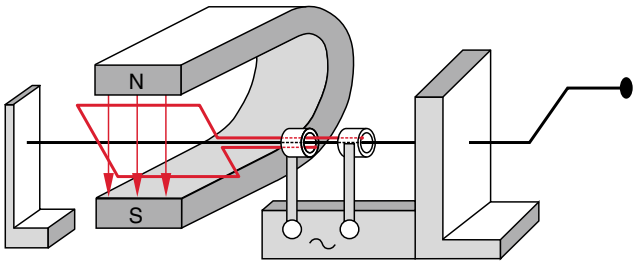


Fig. 4.5 Model of a hand-driven ac-generator. The outside of the sliding contacts are conductors, their inner sides isolating

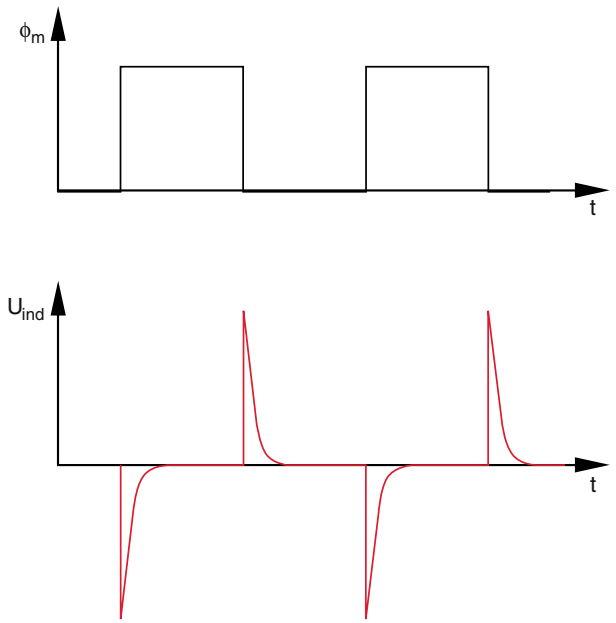


Fig. 4.6 For a rectangular modulation of the magnetic field sharp peaks of the induced voltage appear between the ends of the fixed coil which are $U_{ind} = -d\Phi_m/dt$

2. If we apply a rectangular current to a field coil, then also the magnetic flux through a test coil is modulated nearly rectangular (Fig. 4.6). The time dependence of the induced voltage shows peaks at the rise and drop of the current. The peaks have the opposite sign as the rise and fall of the rectangular current. From (4.2) the change of the magnetic flux follows by integration over time

$$\Delta\Phi_m = \int d\Phi_m = -\int U_{ind} dt. \tag{4.2c}$$

The integral $\int U_{ind} \cdot dt$ gives the area under the curve U_{ind} and is a measure of the change $\Delta\Phi$ of the magnetic flux within the time interval $\Delta t = t_2 - t_1$.

Now we consider a coil with only one winding, which encloses the area A. If the magnetic field changes while the cross section area of the coil and the direction of the magnetic field are kept constant an electric voltage is generated across the ends of the test coil

$$U_{ind} = -\int \dot{\boldsymbol{B}} \cdot d\boldsymbol{A}. \tag{4.2d}$$

This voltage can be traced back to an electric field $\boldsymbol{E}$, because according to (1.13) is

$$U = \int \boldsymbol{E} \cdot ds,$$

where the integration is performed over the circumference of the conductor loop. With Stokes' theorem is

$$\int \boldsymbol{E} \cdot ds = \int \mathbf{rot}\, \boldsymbol{E} \cdot d\boldsymbol{A}. \tag{4.3}$$

Since this is valid for arbitrary surfaces it follows from the comparison of (4.2a) and (4.3)

$$\mathbf{rot}\, \boldsymbol{E} = -\frac{d\boldsymbol{B}}{dt}. \tag{4.4}$$

In words:

A temporally changing magnetic field creates an electric eddy field.

Note The electric field created by charges (Fig. 4.7a) is conservative and **rot** $\boldsymbol{E} = \mathbf{0}$. Therefore, $\boldsymbol{E}$ can be written as a gradient of an electric potential: $\boldsymbol{E} = -\mathbf{grad}\, \phi_{el}$. The electric field lines start on positive charges and end on negative ones. They are not closed. In contrast to this constant electric field generated by static charges where **rot** $\boldsymbol{E} = 0$, is **rot** $\boldsymbol{E} \neq 0$ for that part of the field $\boldsymbol{E}$ created by a temporary changing

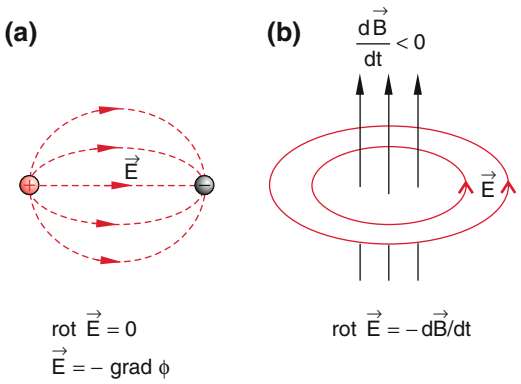


Fig. 4.7 The two sources of electric fields are: **a**) stationary electric charges which generate a rotation-free stationary field, **b**) a changing magnetic field, which produce an electric field with closed field lines. For $d\boldsymbol{B}/dt > 0$ the direction of $\boldsymbol{E}$ reverses

magnetic field (Fig. 4.7b). For this field the electric field lines are closed, and the electric field cannot be written as the gradient of a scalar potential.

4.2 Lenz's Rule

From the negative sign in the induction law (4.2) the following fact can be concluded known as *Lenz' rule.*

- The change of the induction voltage U_{ind} is opposite to the change of the magnetic flux. The electric currents caused in a circuit by the induction voltage generate a magnetic field with a sign which depends on the sign of the magnetic flux Φ_m. It points into the direction of the initial field $\boldsymbol{B}_0$, if $d\boldsymbol{B}_0/dt < 0$ but points into the opposite direction if $d\boldsymbol{B}_0/dt > 0$. The change $d\boldsymbol{B}_0/dt$ of the original field $\boldsymbol{B}_0$ is therefore reduced by the induced magnetic field.
- The direction of the currents induced by the motion of a conductor in a magnetic field is such, that the induced currents impede the motion which generates the currents.

This can be generalized as follows: The currents, fields, and forces created by induction always hinder the process that has initiated the induction (*Lenz's rule*).

This is illustrated in Fig. 4.8.

Lenz's rule should be further illustrated by some experimental examples.

4.2.1 Motion Initiated by Induction

If the north pole of a bar magnet is moved towards an aluminum ring suspended as a pendulum (Fig. 4.9), the direction

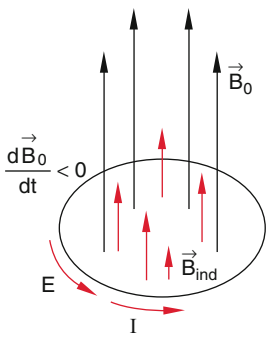


Fig. 4.8 The direction of current I, electric field $\boldsymbol{E}$ and induced magnetic field $\boldsymbol{B}_{ind}$, when $\boldsymbol{B}$ decreases in time. When $\boldsymbol{B}$ increases all red arrows are reversed

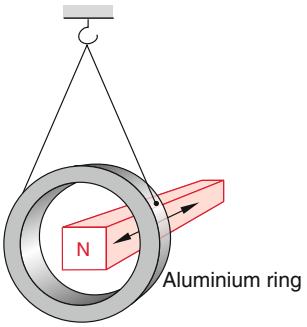


Fig. 4.9 Experimental demonstration of Lenz' rule. The aluminum ring is always repelled when the magnetic rod approaches, it is attracted when up the market rod is withdrawn. This is independent of the fact whether he south- or the north-pole is closer to the ring

of the current induced in the ring is such that the north pole of the induced magnetic dipole points against the north pole of the permanent magnet and therefore, the ring is pushed off by the magnet. Of course, also the magnet is pushed off by the ring and the motion of the magnet towards the ring is hindered.

Now, if we pull back the bar magnet, the induced current and the dipole invert their direction and the movement of the magnet is hindered again. A periodic repetition of these movements can force the ring to swing. We can explain this behavior by considering the conservation of energy.

While the bar magnet approaches the ring, work has to be done that is used to build up the magnetic field created by the induced current flowing in the ring. This magnetic energy is converted to mechanical energy of the moving ring that has been lifted from its rest position.

If we repeat the experiment with an identical ring but with a slit which prevents a circular current through the ring, we observe no motion of the ring because no induced currents can arise.

4.2.2 Electromagnetic Catapult

A field coil with its axis vertically oriented is filled with a bar of soft magnetic iron (Fig. 4.10). Above the coil lies a thick ring of aluminum (or copper). When switching on a current through the field coil the increasing magnetic field induces a current in the ring above the coil. The two magnetic moments of ring and coil are directed so that they repel each other and the ring is thrown upward. Be careful. Heights of up to 10 m are accessible. A technical application of this principle is the acceleration of small particles to high speeds which even can exceed the escape velocity of 11 km/s from the earth surface [1, 2].

4.2.3 Magnetic Levitation

Magnetic levitation means the floating of a thick slab of aluminum in the field of an electromagnet driven by an

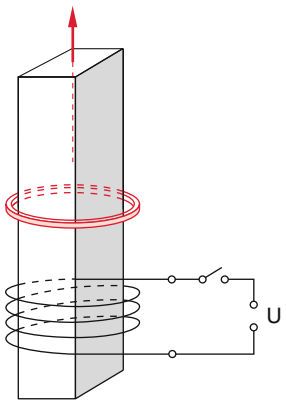


Fig. 4.10 Electro-magnetic induction gun

alternating current. The slab floats a few centimeters above the magnet (Fig. 4.11).

Here, again a varying magnetic flux induces currents and their magnetic moment creates a repulsive force balancing the gravitational force. To avoid the slab getting out of its central position, there is an additional coil about the electromagnet that creates the stabilizing magnetic field. The thicker the slab the higher is the repelling force because of higher induced currents [3].

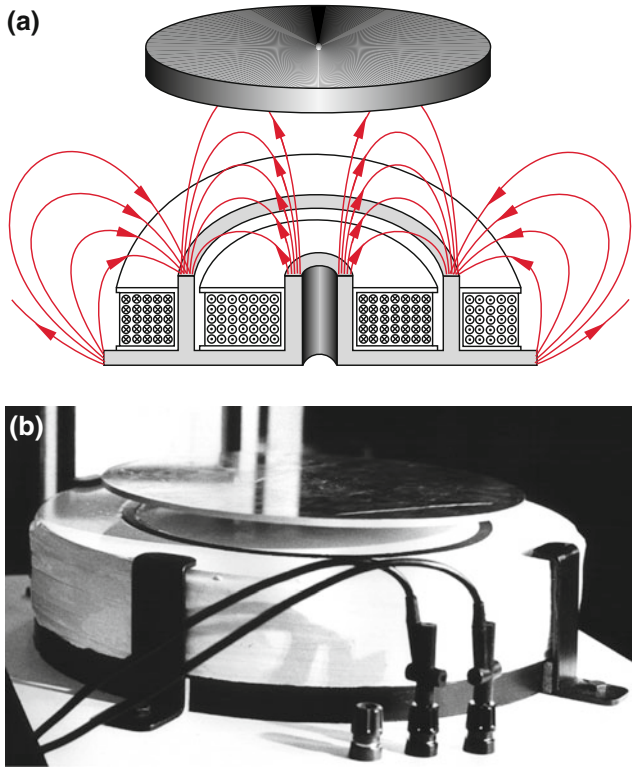


Fig. 4.11 Eddy current levitometer **a**) schematic drawing, **b**) photo-picture of the actual design (with kind permission of Prof. Gruben)

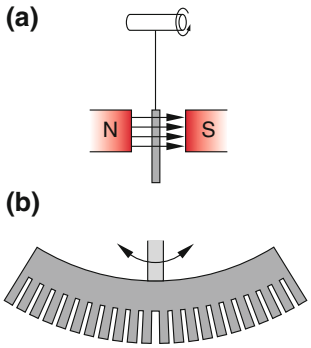


Fig. 4.12 Waltenhof pendulum **a**) side view, **b**) front view

4.2.4 Eddy Currents

Induced currents in an extended conductor are called eddy currents. Their direction and magnitude depend on the temporal variation $d\boldsymbol{B}/dt$ of the surrounding magnetic field and of the spatial distribution of the electric resistance $R(x, y, z)$. These eddy currents and Lenz's law is clearly demonstrated by the experiment shown in Fig. 4.12 (Waltenhofen's pendulum).

A solid aluminum disc is fixed at one end of a long stick and is swinging between the poles of a currentless electromagnet. At time $t = 0$ the current through the electromagnet is switched on, the oscillation of the pendulum dies away nearly immediately. The reason for this observation are the high eddy currents in the disc that are dissipated by Joule's heating. The mechanical energy of the swinging pendulum is transferred into heat.

Sawing many gaps into the disc, perpendicular to the direction of motion (Fig. 4.12b) only lower eddy currents can be induced and the damping of the pendulum is correspondingly smaller.

The eddy current brake is applied in many electrically driven vehicles for emergency breaking.

4.3 Self Inductance and Mutual Inductance

For technical application of coils or other arrangements of conductors it would be very annoying to calculate the integral (4.2) every time anew. To simplify the calculations we assign a scalar quantity, called the inductance, to every arbitrary arrangement of conductors. However, the calculation of the inductance itself can be for some arrangements complicated and therefore its experimental determination may be often the better choice.

4.3.1 Self Inductance

The magnetic flux through a coil changes with the temporal change of the current through the coil. Therefore, according to Faraday's induction law, even in the coil itself an induction voltage is generated. Due to Lenz's law the sign of this voltage is opposite to the external voltage driving the current through the coil.

Since the magnetic field of the coil is proportional to the current I through the coil it follows for the magnetic flux

$$\Phi_m = \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{F} = L \cdot I, \tag{4.5a}$$

The proportionality constant L with the unit

$$[L] = 1\,\mathrm{V\,s/A} = 1\,\mathrm{Henry} = 1\,\mathrm{H}$$

is the ***coefficient of self-induction***. The induced voltage is then obtained from (4.2) as

$$U_{\mathrm{ind}} = -L \cdot \frac{\mathrm{d}I}{\mathrm{d}t}. \tag{4.5b}$$

4.3.1.1 Switching on the Supply Voltage

In the circuit of Fig. 4.13a the external constant voltage U_0 is supplied to the circuit by closing the switch S at time $t = 0$. According to Kirchhoff's rule (see Sect. 2.6) we obtain the relation

$$U_0 = I \cdot R - U_{\mathrm{ind}} = I \cdot R + L \cdot \frac{\mathrm{d}I}{\mathrm{d}t}, \tag{4.6}$$

where R is the Ohmic resistance of the coil. With the ansatz

$$I(t) = K \cdot \mathrm{e}^{-(R/L)\cdot t} + I_0 \tag{4.7a}$$

The solution of the inhomogeneous differential equation becomes with the initial condition $I(0) = 0$

$$I(t) = \frac{U_0}{R} \cdot (1 - \mathrm{e}^{-(R/L)\cdot t}). \tag{4.7b}$$

The current does not reach immediately at $t = 0$ its final value $I = U_0/R$, expected by Ohm's law, but increases gradually to the final value. The time delay depends on the ratio of self inductance L and Ohmic resistance R. This ratio $\tau = L/R$ is called the time constant of the circuit. At the time $t = \tau$ the current $I(t)$ has reached the value $I(\tau) = U_0/R \cdot (1 - 1/\mathrm{e}) \sim 0.63 \cdot I(\infty)$.

This time dependence $I(t)$ can be directly viewed on an oscilloscope. A more qualitative visual demonstration is possible with the circuit shown in Fig. 4.13c, which includes the two light bulbs G_1 and G_2. When the switch S is closed, at first the lamp G_1 lights up and only after the time $\tau = L/R$ the light bulb G_2 shines. With sufficient large values of L the delay time τ can increase to several seconds. After the stationary state has been reached both lamps shine equally bright, because the same current I flows through both lamps if the resistances $R_1 = R_2$ are equal.

4.3.1.2 Switching off the Supply Voltage

Now we consider the opposite process: the switch S in Fig. 4.14 is closed for $t = 0$ and the current I_1 through the resistor R_1 is $I_1(t < 0) = U_0/R_1$ while the current through the coil is $I_2(t < 0) = U_0/R_2$. For $R_1 > R_2$ is $I_1 < I_2$. At $t = 0$ the switch S is suddenly opened. With the initial conditions $U_0(t = 0) = 0$ and $I_2(t = 0) = I_0$ we obtain the equation

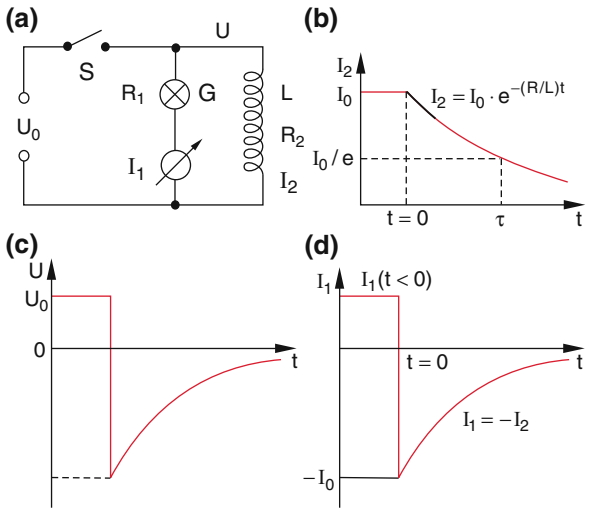


Fig. 4.14 Induction voltage after switching off the current source by opening the switch S **a**) circuit, **b**) current I_2(t), **c**) voltage $U(t)$, **d**) current $I_1(t)$

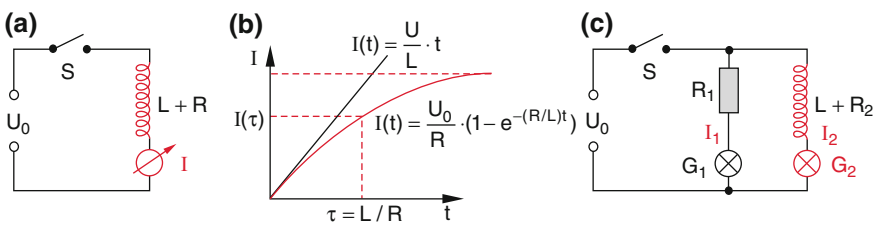


Fig. 4.13 Demonstration of sel- induction of a coil **a**) experimental setup, **b**) electric current $I(t)$ after closing the switch S, **c**) illustration of the delay of the current $I(t)$ by two small light bulbs G_1 and G_2

$$0 = I_2 \cdot R - U_{\text{ind}} = I_2 \cdot R + L \cdot \frac{\mathrm{d}I_2}{\mathrm{d}t} \tag{4.8a}$$

With the solution

$$I_2(t) = I_0 \cdot \mathrm{e}^{-(R/L)\cdot t} \tag{4.8b}$$

With $R = R_1 + R_2$. Here R_1 is a pure Ohmic resistor while R_2 is the Ohmic resistance of the coil with inductance L. After opening the switch S at $t = 0$ the current I_2 does not jump immediately to zero but decreases exponentially with the time constant $\tau = L/R$. Across the coil the induction voltage

$$U_{\text{ind}} = -I_2(R_1 + R_2) = -L\frac{\mathrm{d}I_2}{\mathrm{d}t} \tag{4.8c}$$

appears (Fig. 4.14c) and through the ampere meter the current $I_1 = -I_2$ flows, which is opposite but larger than $I_1(t < 0)$. With $U_0 = I_0 \cdot R_2$ we get the induced voltage

$$U_{\text{ind}} = -U_0 \frac{R_1 + R_2}{R_2} \mathrm{e}^{-(R/L)t}, \tag{4.8d}$$

This shows that for $R_1 \gg R_2$ the induced voltage U_{ind} $(t = 0) \approx (R_1/R_2) \cdot U_0$ is essentially higher than U_0. Therefore also the current $I = U_{\text{ind}}/R_0$ through R_1 is much higher than before the opening of the switch S. If, for instance R_1 is replaced by a light bulb it will suddenly flash very brightly and may even blow up if L is sufficiently large.

4.3.1.3 Ignition of Fluorescent Tubes

Fluorescent tubes are long evacuated glass tubes which are filled with a noble gas and a small addition of mercury. At both ends of the tube are heated filaments. The inner wall of the glass tube is covered with a thin film of fluorescent material which converts the ultraviolet light of the mercury-noble gas mixture into a continuous spectrum in the visible range. This makes the total spectrum of the fluorescent tube similar to daylight.

The ignition is explained in Fig. 4.15. Closing the switch S connects the tube to the power supply. Now the line voltage appears between the two filaments and across the starter, which consists of a small glow discharge lamp where one electrode is a bimetallic strip. When the discharge ignites, the bimetallic strip bends and closes a contact. This causes a short cut across the discharge of the starter which therefore goes off.

Now the full current flows through the filaments of the fluorescent tube, which heat up and evaporate the mercury deposited on the filaments. Furthermore the hot filaments emit electrons which can ionize the gas atoms in the tube.

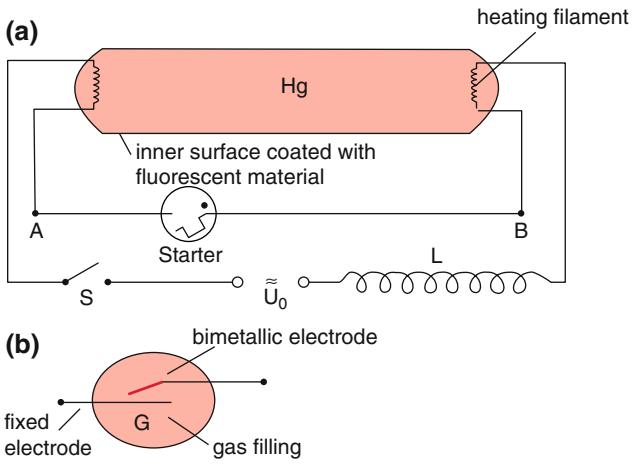


Fig. 4.15 Discharge lamp, **a**) electric circuit, **b**) ignition device (starter) as glow discharge in a gas-filled bulb wit a bimetallic switch

When the discharge in the starter has gone off, the bimetallic strip bends back and opens the circuit again. This sudden interruption of the current I through the coil with inductivity L causes a high induction voltage $U_{\text{ind}} = L \cdot \mathrm{d}I/\mathrm{d}t$ (about 1 kV) across the coil, which appears between the filaments of the tube. This causes the ignition of the discharge in the tube. The interruption of the current before the ignition is so short that the filaments do not cool down during this short time. Since the gas discharge is now ignited, the electron and ion bombardment of the filaments leads to a continuous heating of the filaments, which therefore continue to emit electrons and support the discharge, The voltage between the point A and B in Fig. 4.15 drops to the much lower burning voltage U_B (100–120 V). The difference $\Delta U = U_0 - U_B$ to the supply voltage U_0 (in Germany 230 V) appears across the inductance L. For an ac-current with frequency ω this is $\Delta U = \omega \cdot L$ (see Sect. 5.4).

The coil L has a double function: It delivers the high ignition voltage to start the discharge in the fluorescent tube and it limits the discharge current through the tube by providing a dropping resistor $R = \omega \cdot L$ [4].

Instead of mercury one can also use sodium, where the light emission efficiency is very high in the yellow spectral region. Such lamps therefore emit yellow light.

In Fig. 4.16 a commercial starter with bimetallic switch is illustrated. Modern designs use instead of the bimetallic switch electronic switches, which have the additional advantage, that the ignition time can be chosen at the peak of the ac supply voltage, which makes the ignition process more reliable. The fluorescent tube reaches its maximum brightness only after some minutes, when the stationary vapor pressure of the mercury has been reached.

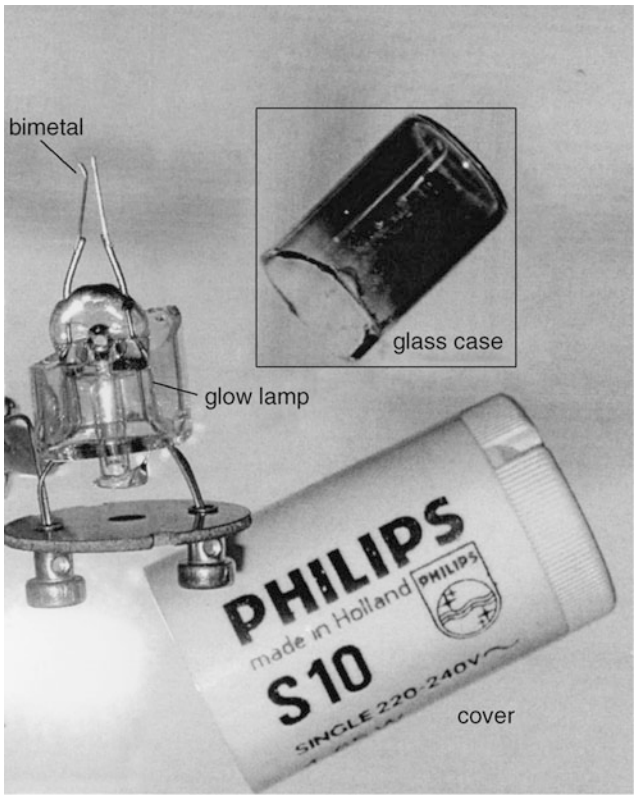


Fig. 4.16 Ignition device with bimetallic switch for a discharge lamp

4.3.1.4 Self-inductance Coefficient of a Solenoid

The magnetic field in a solenoid with length l and n windings per m which carries a current I (Fig. 4.17) is, according to (3.10)

$$B = \mu_0 \cdot n \cdot I.$$

The magnetic flux through one winding of the solenoid with cross section A is

$$\Phi_m = B \cdot A = \mu_0 \cdot n \cdot I \cdot A$$

The temporal change of Φ_m due to a change dI/dt of the current is

$$\frac{d\Phi_m}{dt} = \mu_0 \cdot n \cdot A \cdot \frac{dI}{dt}.$$

Between the ends of the solenoid with $N = n \cdot l$ windings the voltage

$$U_{ind} = -N \cdot \dot{\Phi}_m = -\mu_0 n^2 l A \cdot \frac{dI}{dt} = -L \cdot \frac{dI}{dt} \tag{4.9}$$

is induced. The self-inductance coefficient L is therefore

$$L = \mu_0 \cdot n^2 \cdot V, \tag{4.10}$$

here $V = lA$ is the volume that is included by the solenoid.

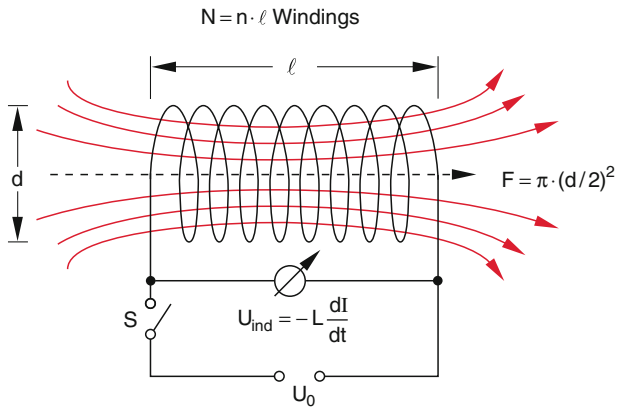


Fig. 4.17 Self-inductance of a solenoid

4.3.1.5 Self-induction of a Double Circuit Line

Two long parallel wires with radius r_0 and distance d which carry the current I in opposite directions are called a double circuit line (Fig. 4.18). It represents an important element for the transmission of electric power.

When the direction of the wires is the z-direction, the magnetic field $\boldsymbol{B}$ lies in the x-y-plane. On the connecting line between the two wires, which we chose as the x-direction, the amount of $\boldsymbol{B}$ outside the two wires is:

$$B^{(out)} = \frac{\mu_0 I}{2\pi}\left(\frac{1}{\frac{d}{2}+x} + \frac{1}{\frac{d}{2}-x}\right). \tag{4.11}$$

Between the two lines is B for the left part of the spacing $d(I > 0,\ x < 0)$

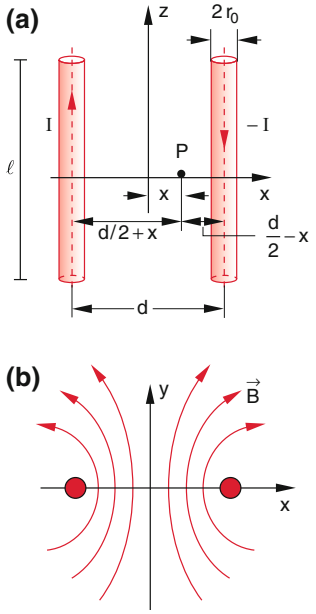


Fig. 4.18 Parallel double line **a**) arrangement, **b**) magnetic field in the x-y- plane

$$B_1^{(\mathrm{in})} = \frac{\mu_0 I}{2\pi r_0^2}\left(\frac{d}{2} + x\right) \tag{4.12a}$$

And for the right part of the spacing:

$$B_\mathrm{r}^{(\mathrm{i})} = \frac{\mu_0 I}{2\pi r_0^2}\left(\frac{d}{2} - x\right). \tag{4.12b}$$

The magnetic flux Φ_m through a part of the double circuit line with length l and cross section area $A = d \cdot l$ in the x-y-plane is then

$$\begin{aligned}\Phi_m &= l \cdot \left[\int_{-d/2+r_0}^{d/2-r_0} B^{(\mathrm{out})}\mathrm{d}x + \int_{-d/2}^{-d/2+r_0} B_1^{(\mathrm{in})}\mathrm{d}x + \int_{d/2-r_0}^{d/2} B_\mathrm{r}^{(\mathrm{i})}\mathrm{d}x\right] \\ &= \frac{\mu_0 \cdot I \cdot l}{\pi} \cdot \left[\frac{1}{2} + \ln\frac{d - r_0}{r_0}\right].\end{aligned} \tag{4.13a}$$

The self-induction coefficient becomes

$$L = \frac{\Phi_m}{I} = \frac{\mu_0 \cdot l}{\pi} \cdot \left[\frac{1}{2} + \ln\frac{d - r_0}{r_0}\right]. \tag{4.13b}$$

Equation (4.13b) shows that the self-induction coefficient L of a double circuit line increases logarithmically with the distance d between the two wires.

Note, that L increases with decreasing radius r_0 of the wires. Therefore flat ribbons are used for double circuit lines with low inductance which are separated by a thin insulating layer. For $d = 2r_0$ one obtains from (4.13b) the minimum inductance

$$L(d = 2r_0) = \frac{\mu_0 \cdot l}{2\pi}. \tag{4.13c}$$

4.3.2 Mutual Induction

We consider a circuit carrying the current I_1 (Fig. 4.19). According to the *Biot-Savart-law* (see Sect. 3.2.5) this circuit generates in a point $P(\boldsymbol{r}_2)$ a magnetic field $\boldsymbol{B}$ with the vector potential

$$\boldsymbol{A}(\boldsymbol{r}_2) = \frac{\mu_0 I_1}{4\pi}\int_{s_1}\frac{\mathrm{d}\boldsymbol{s}_1}{r_{12}}, \tag{4.14}$$

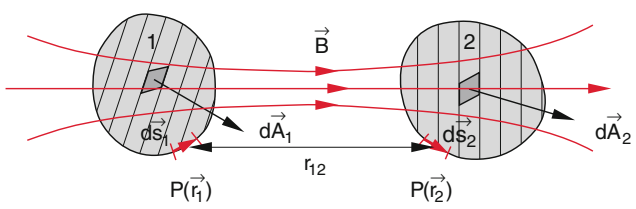


Fig. 4.19 Illustration of the coefficient L_{12} of mutual inductance

where $\mathrm{d}\boldsymbol{s}_1$ is a line element of the circuit. This magnetic field causes a magnetic flux

$$\begin{aligned}\Phi_m &= \int_F \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A} \\ &= \int_F \mathbf{rot}\,\boldsymbol{A} \cdot \mathrm{d}\boldsymbol{A} = \int_{s_2} \boldsymbol{A} \cdot \mathrm{d}s_2\end{aligned} \tag{4.15}$$

through the area A which is encircled by a second conductor loop with the line element $\mathrm{d}\boldsymbol{s}_2$.

Note Unfortunately the vector potential $\boldsymbol{A}$ and the area $\boldsymbol{A}$ are labeled with the same letter. The careful reader would not be confused by this convention.

The last equality follows from Stokes' law (see textbooks on vector analysis).

Inserting (4.14) into (4.15) one obtains the magnetic flux through the second loop, which causes the current I_2

$$\Phi_m = \frac{\mu_0 I_1}{4\pi}\int_{s_1}\int_{s_2}\frac{\mathrm{d}\boldsymbol{s}_1 \cdot \mathrm{d}\boldsymbol{s}_2}{r_{12}} = L_{12} \cdot I_2. \tag{4.16}$$

The proportionality factor

$$L_{12} = L_{21} = \frac{\mu_0}{4\pi}\int_{s_1}\int_{s_2}\frac{\mathrm{d}\boldsymbol{s}_1 \cdot \mathrm{d}\boldsymbol{s}_2}{r_{12}} \tag{4.17}$$

Is the **coefficient of mutual inductance**. It depends on the geometric shape of the two circuits, from their mutual orientation and their distance.

For arbitrary configurations (4.17) is generally only numerically solvable. We will therefore illustrate (4.17) for some simple examples.

4.3.2.1 Rectangular Conductor Loop in a Homogeneous Magnetic Field

We consider a rectangular loop which encloses the area A in a homogeneous magnetic field of a solenoid with current I and n windings per meter (Fig. 4.3). The magnetic flux through the loop is according to (3.10)

$$\Phi_m = \int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{A} = \mu_0 \cdot n \cdot I \cdot A \cdot \cos\alpha,$$

where α is the angel between the surface normal $\boldsymbol{A}$ and the axis of the solenoid. The coefficient of mutual inductance is then

$$L_{12} = \mu_0 \cdot n \cdot A \cdot \cos\alpha.$$

It becomes zero for $\alpha = 90°$.

4.3.2.2 Two Circular Loops with Different Relative Orientation

In Fig. 4.20 two circular loops with radius R are shown with different orientations. The maximum value of L_{12} is obtained for $\alpha = 0°$, i.e. both loops are parallel (Fig. 4.20b). The vertical arrangement of Fig. 4.20c has the smallest value $L_{12} = 0$, because the magnetic field of the first loop is parallel to the plane of the second loop and the magnetic flux through the second loop is zero.

For the parallel arrangement in Fig. 4.20b and for small distances $d \ll R$ nearly the whole magnetic flux generated by the first loop passes through the second loop. Therefor, according to (3.19) the coefficient

$$L_{12} = \frac{\pi}{2}\mu_0 R \tag{4.18a}$$

is for $d \ll R$ independent of the distance d.

For large distances ($d \gg R$) is according to (3.20)

$$B \approx \frac{\mu_0}{2\pi} \cdot \frac{I \cdot A}{d^3}, \tag{4.18b}$$

The coefficient of mutual inductance is then

$$L_{12} \approx \frac{\pi}{2}\mu_0 \cdot \frac{R^4}{d^3} \tag{4.18c}$$

While for the two limiting cases ($d \ll R$ and $d \gg R$) the determination of L_{12} is relative easy, for the general case the integral (4.17) has to be calculated. This leads to elliptical integrals, where the solutions are only approximately possible. If an iron bar is arranged through both loops the magnetic flux Φ_m through the loop 2 becomes larger, because the magnetic field, generated by loop 1 is amplified by the iron bar and is guided through the bar into loop 2 (Fig. 4.20d).

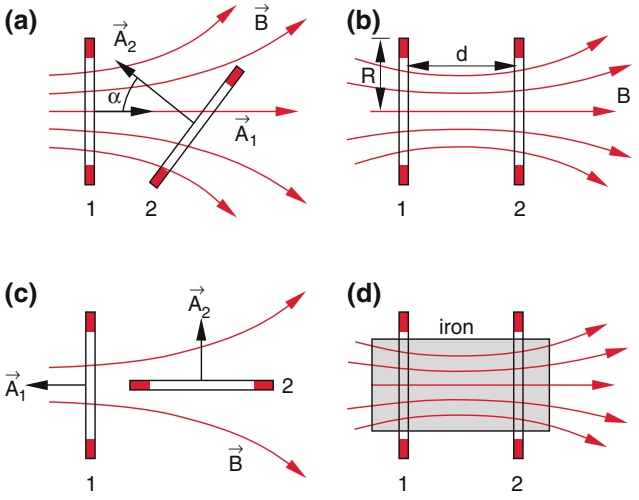


Fig. 4.20 Coefficient L_{12} of mutual inductance for two circular loops with equal but different mutual orientations

Using ferromagnetic materials with a large relative permeability μ the coupling between two conductive loops can be very much enhanced, which enlarges the coefficient L_{12}. This is utilized in electric transformers (see Sect. 5.6).

4.4 The Energy of the Magnetic Field

The energy dissipated in the resistor R in Fig. 4.14 after the disconnection of the voltage supply must have come from the magnetic field of the solenoid. The energy of the magnetic field is therefore

$$W_{magn} = \int_0^\infty I \cdot U \cdot dt = \int_0^\infty I^2 \cdot R \cdot dt. \tag{4.19a}$$

With (4.8a, 4.8b, 4.8c, 4.8d) this gives

$$\begin{aligned} W_{magn} &= \int_0^\infty I_0^2 \cdot e^{-(2R/L)\cdot t} \cdot R \cdot dt \\ &= \frac{1}{2} I_0^2 \cdot L, \end{aligned} \tag{4.19b}$$

where $I_0 = I(t < 0)$ is the stationary current through the solenoid bevor the disconnection from the source.

Magnetic fields can be therefore used as energy storage. If they are generated by currents in superconducting coils one does not need any power to maintain the stationary magnetic field (apart from the power to cool the system down below the transition temperature to superconductivity.

With $L = \mu_0 \cdot n^2 \cdot V$ (see (4.10)) we obtain the energy density of the magnetic field

$$w_{magn} = \frac{W_{magn}}{V} = \frac{1}{2}\mu_0 \cdot n^2 \cdot I_0^2 = \frac{B^2}{2\mu_0}. \tag{4.19c}$$

Remark Compare the corresponding expressions for energy W and energy density w of the electric and the magnetic field.

$$\begin{aligned} W_{el} &= \frac{1}{2} C U^2 \\ W_{magn} &= \frac{1}{2} L I^2 \\ w_{el} &= \frac{1}{2}\varepsilon_0 E^2 \\ w_{magn} &= \frac{1}{2}\mu_0 H^2 = \frac{1}{2\mu_0} B^2. \end{aligned} \tag{4.19d}$$

Using the relation $\varepsilon_0 \cdot \mu_0 = 1/c^2$ we can write the energy density of the electromagnetic field in vacuum as

$$w_{\text{em}} = \frac{1}{2}\varepsilon_0(E^2 + c^2B^2). \tag{4.20a}$$

The energy density of the electromagnetic field in matter with the relative permittivity ε and the relative permeability μ is

$$w_{\text{em}} = \frac{1}{2}\varepsilon_0\left(\varepsilon E^2 + \frac{c^2}{\mu}B^2\right) \tag{4.20b}$$

With $D = \varepsilon \cdot \varepsilon_0 \cdot E$ and $H = B/(\mu \cdot \mu_0)$ we can write (4.20c) as

$$w_{\text{em}} = \frac{1}{2}(E \cdot D + B \cdot H). \tag{4.20c}$$

4.5 The Displacement Current

In many cases the formulation of Ampere's law

$$\oint \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{s} = \mu_0 I = \mu_0 \int_F j \cdot \mathrm{d}\boldsymbol{A} \tag{4.21a}$$

used in Chap. 3 is not unambiguous. If the differential form (3.7)

$$\text{rot}\, B = \mu_0 \cdot j$$

shall be derived form (3.6) the Eq. (3.6) must be valid for arbitrary paths around the conductor and for arbitrary areas which are encircled by these paths.

In Fig. 4.21b an electric circuit is shown with a capacitor C that carries a time-varying current $I(t)$. If the circular curve s_1 is chosen as integration path the circular area dA can be inserted into the integral in (4.21a) as well as any area circumvented by an arbitrary curve s_1. If the curve s_2 inside the capacitor C is chosen, the current density j, defined in the conventional way, is zero. The magnetic field measured at the point P_1 is given by Eq. (3.6) with the integration path s_1. Choosing the path s_2 inside the capacitor, the magnetic field would be zero. However, the whole current circuit, including the capacitor carries the alternating current I. What is wrong with the conventional definition?

In order to remove the discrepancy *Clark Maxwell* (1831–1879, Fig. 4.21a) introduced the term "**displacement current**" with the following argument: If an ac-current flows through the conducting line in Fig. 4.21, the charge Q on the capacitor plates changes. This leads to a change of the electric field between the plates. With the relation

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(\varepsilon_0 \boldsymbol{A} \cdot \boldsymbol{E}) = \varepsilon_0 \boldsymbol{A} \cdot \frac{\partial \boldsymbol{E}}{\partial t} \tag{4.21b}$$

between the charge $Q = \varepsilon_0 \cdot A \cdot E$ on the plates with area A and the electric field E inside the capacitor a current $I_V = \varepsilon_0 \cdot A \cdot \partial E/\partial t$ can be defined which is called "*displacement current*". The corresponding current density is then

$$j_D = \varepsilon_0 \cdot \frac{\partial \boldsymbol{E}}{\partial t} \tag{4.22}$$

which is proportional to the time change $\partial E/\partial t$ of the electric field inside the capacitor. Here the partial derivative is chosen, because (4.22) is also valid for inhomogeneous fields where $\boldsymbol{E}(\boldsymbol{r}, t)$ also depends on the spatial coordinates. When (4.22) is added to the current density $\boldsymbol{j}$ the total current density is then $\boldsymbol{j} + \boldsymbol{j}_D$. Inserting this into (3.6) we get

$$\int \boldsymbol{B} \cdot \mathrm{d}s = \mu_0 I = \mu_0 \int (\boldsymbol{j} + \boldsymbol{j}_D) \cdot \mathrm{d}\boldsymbol{A} \tag{4.23a}$$

or in the differential form (3.7)

$$\begin{aligned} \text{rot}\, \boldsymbol{B} &= \mu_0(\boldsymbol{j} + \boldsymbol{j}_D) \\ &= \mu_0 \boldsymbol{j} + \mu_0\varepsilon_0 \frac{\partial \boldsymbol{E}}{\partial t}. \end{aligned} \tag{4.23b}$$

With $\mu_0 \cdot \varepsilon_0 = 1/c^2$ we can write (4.23b) as

$$\text{rot}\, \boldsymbol{B} = \mu_0 \boldsymbol{j} + \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t}. \tag{4.23c}$$

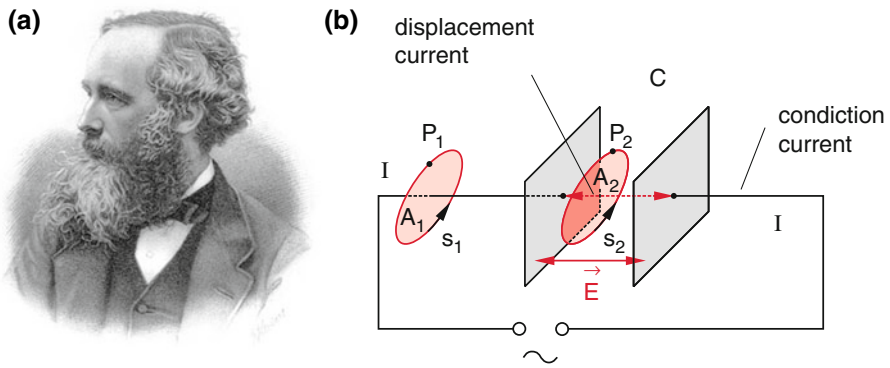


Fig. 4.21 **a**) James Clerk Maxwell **b**) Illustration of the displacement current

This important result implies:

Magnetic fields are not only generated by electric currents, but also by time-varying electric fields.

Without this mechanism there would be no electromagnetic waves and we would not receive any sun light.

Remark Due to the introduction of the displacement current the continuity equation is fulfilled. This implies that the preservation of charge is saved. This would not be the case without the term $\partial E/\partial t$ in (4.23c), as can be seen by the following arguments:

When we apply the operator div to Eq. (4.23c) we obtain:

$$\text{div } \textbf{rot } \boldsymbol{B} = \mu_0 \text{div} \boldsymbol{j} + \varepsilon_0 \mu_0 \frac{\partial}{\partial t} \text{div } \boldsymbol{E} = 0 \qquad (4.23\text{d})$$

because div **rot** $\boldsymbol{B} = \nabla \cdot (\nabla \times \boldsymbol{B}) = 0$ since the scalar product of two orthogonal vectors ∇ and $(\nabla \times \boldsymbol{B})$ is zero.

Inserting in (4.23d) $\text{div } \boldsymbol{E} = \varrho/\varepsilon_0$ we get the continuity equation

$$\text{div} \boldsymbol{j} + \frac{\partial \varrho}{\partial t} = 0 \qquad (4.23\text{e})$$

Equation (4.23a, 4.23b) can be checked experimentally by applying a high frequency voltage

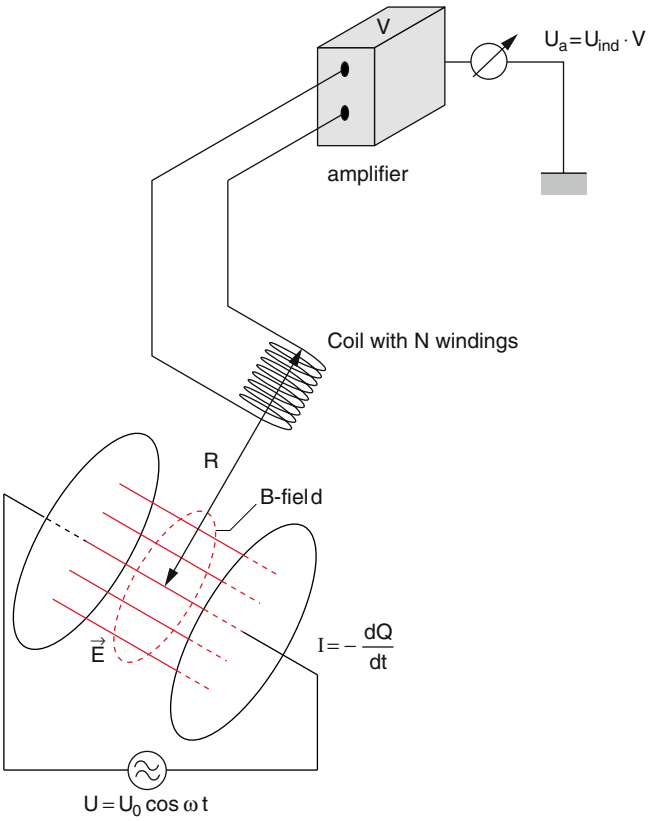


Fig. 4.22 Experimental proof of Eqs. (4.23a–4.23h)

$$U_C = U_0 \cdot \cos \omega t$$

to the arrangement in Fig. 4.21. The displacement current is then

$$I_D = \frac{dQ}{dt} = C \cdot \frac{dU_C}{dt} = -C \cdot U_0 \cdot \omega \cdot \sin \omega t. \qquad (4.23\text{f})$$

where C is the capacity of the capacitor. In case of circular capacitor plates the magnetic field lines are circles around the symmetry axis of the capacitor in the planes $x = \text{const}$ (Fig. 4.22).

According to Eq. (3.8) the magnetic field B at the edge of the capacitor at the distance R_0 from the symmetry axis is

$$B = \frac{\mu_0 \cdot I_D}{(2\pi R_0)}.$$

The magnetic flux Φ_m through a small induction coil with N windings and its surface vector $\boldsymbol{A}$ normal to the area A of the coil and parallel to the magnetic field is

$$\Phi_m = N \cdot A \cdot B$$

The voltage induced in the coil by the alternating magnetic field is

$$U_{ind} = -N \cdot A \cdot \frac{dB}{dt} = -\frac{\mu_0}{2\pi R_0} N \cdot A \cdot C \cdot \frac{d^2 U_C}{dt^2} \qquad (4.23\text{g})$$

With the amplitude

$$U_{ind}^{max} = \frac{\mu_0}{2\pi R_0} N \cdot A \cdot C \cdot U_0 \cdot \omega^2. \qquad (4.23\text{h})$$

Example

$A = 10^{-4}$ m^2; $N = 10^3$; $R_0 = 0.2$ m; $U_0 = 100$ V; $\omega = 2\pi \cdot 10^6$ s^{-1}; $d = 0.1$ m, $\Longrightarrow C = \varepsilon_0 \cdot \pi R_0^2/d = 11 \cdot 10^{-12}$ F $\Longrightarrow U_{ind}^{max} = (4\pi \cdot 10^{-7})/(2\pi \cdot 0.2) \cdot 11 \cdot 10^{-12} \cdot 10^2 \cdot (2\pi \cdot 10^6)^2$ V = 4.8 mV.

This voltage can be viewed directly on an oscilloscope without any amplifier.

When a dielectricum with the relative permittivity ε and the relative permeability μ is inserted into the capacitor we must modify (4.23c). Instead of the electric field $\boldsymbol{E}$ we have to use the dielectric displacement density $\boldsymbol{D} = \varepsilon_0 \cdot \boldsymbol{E}$. A similar approach is necessary for the magnetic field, when magnetic materials with the permeability constant μ are introduced into the field. Equation (4.23a–4.23h) can be formulated independently of the medium, if the magnetic intensity $\boldsymbol{H}$ is used instead of the magnetic field strength $\boldsymbol{B}$. Instead of (4.23a–4.23h) we get

$$\textbf{rot } \boldsymbol{H} = \boldsymbol{j} + \frac{\partial \boldsymbol{D}}{\partial t}. \qquad (4.24)$$

4.6 Maxwell's Equations and Electrodynamic Potentials

The introduction of the displacement current and with Faraday's law of induction we can extend the field Eqs. (4.1) for stationary charges and currents to temporally varying conditions. Using (4.4) and (4.23c) we arrive at Maxwell's equations

$$\mathbf{rot}\,\boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{4.25a}$$

$$\mathbf{rot}\,\boldsymbol{B} = \mu_0 j + \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t}, \tag{4.25b}$$

$$\mathrm{div}\,\boldsymbol{E} = \frac{\varrho}{\varepsilon_0}, \tag{4.25c}$$

$$\mathrm{div}\,\boldsymbol{B} = 0. \tag{4.25d}$$

With the Eqs. (1.65) and (4.24) we can generalize these equations and get

$$\mathbf{rot}\,\boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{4.26a}$$

$$\mathbf{rot}\,\boldsymbol{H} = j + \frac{\partial \boldsymbol{D}}{\partial t}, \tag{4.26b}$$

$$\mathrm{div}\,\boldsymbol{D} = \varrho, \tag{4.26c}$$

$$\mathrm{div}\,\boldsymbol{B} = 0. \tag{4.26d}$$

Together with the Lorentz force

$$\boldsymbol{F} = q \cdot (\boldsymbol{E} + \boldsymbol{v} \times \boldsymbol{B}) \tag{4.26e}$$

and Newtons equation of motion $\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$ these equations describe all electromagnetic phenomena observed up to now.

Electric fields are generated by charges as well as by temporary varying magnetic fields. Magnetic fields are generated by electric currents as well as by temporary varying electric fields (Fig. 4.23).

Electric and magnetic fields are closely interlinked and form electromagnetic fields.

For temporary constant fields (4.25a–4.25d) reduces to (4.1)

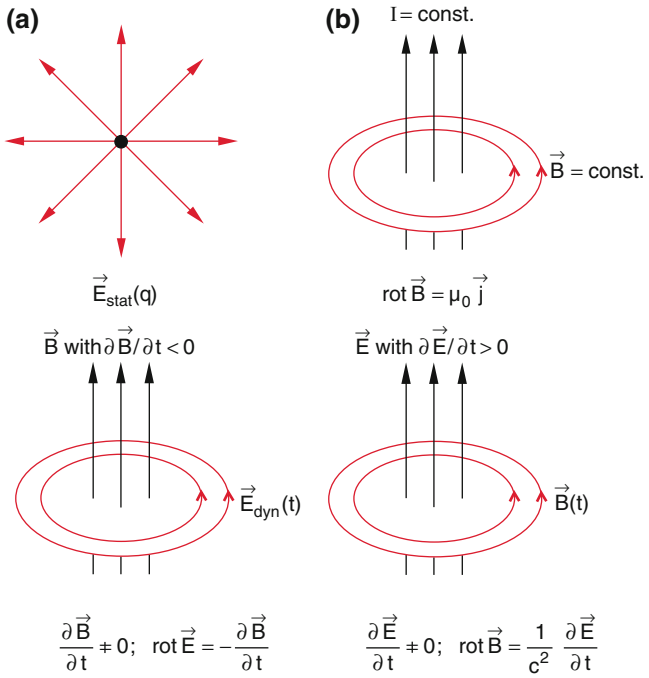


Fig. 4.23 Vivid illustration of the Maxwell equations describing the generation of an electric field by the temporal variation of an magnetic field (Faraday' induction law) and the generation of a magnetic field by time variation of an electric field. Both are compared with static fields

The Maxwell equations, which provide the basis of the whole electrodynamics can be derived from a few general principles:

- The conservation of the electric charge (continuity equation)
- The conservation of the magnetic flux Φ_m (there are no magnetic monopoles)
- The Lorentz force on charges moving in electromagnetic fields.

The proof of these statements would surpass the framework of this introductory textbook and the reader is referred to the corresponding literature [5–7].

The Maxwell Eqs. (4.25a–4.25d) represent a system of coupled differential equations for the fields $\boldsymbol{E}$ and $\boldsymbol{B}$, where the two fields are coupled with each other by the relations (4.25a) and (4.25b).

For the solution of these equations it is often convenient to write the equations in an uncoupled form. This can be achieved by using the *scalar electric potential* ϕ_{el} and the *magnetic vector potential* A with $\mathbf{rot}\,\boldsymbol{A} = \boldsymbol{B}$.

Since $\mathbf{rot}\,\boldsymbol{E} \neq \boldsymbol{0}$ the electric field $\boldsymbol{E}$ can no longer be written as grad ϕ_{el}.

However, we can deduce from (4.25a–4.25d) by interchanging the spatial and the temporal differentiation $\partial B/\partial t = \partial/\partial t(\mathrm{rot}\,A) = \mathrm{rot}(\partial A/\partial t)$ the equation

$$\mathbf{rot}\,\boldsymbol{E} + \partial B/\partial t = \mathbf{rot}\,(\boldsymbol{E} + \partial A/\partial t) = 0, \tag{4.27}$$

This equation allows us to write the sum $\boldsymbol{E} + \partial A/\partial t$ as gradient of a scalar potential

$$\begin{aligned} \boldsymbol{E} + \frac{\partial \boldsymbol{A}}{\partial t} &= -\mathbf{grad}\,\phi_{el} \\ \Rightarrow \boldsymbol{E} &= -\mathbf{grad}\,\phi_{\mathrm{el}} - \frac{\partial \boldsymbol{A}}{\partial t}, \end{aligned} \tag{4.28}$$

For stationary fields ($\partial A/\partial t = 0$) this reduces again to the conventional form $\boldsymbol{E} = -\mathbf{grad}\ \phi_{\mathrm{el}}$ used in electrostatics.

The vector potential $\boldsymbol{A}$ is not unambiguously defined by $\boldsymbol{B} = \mathbf{rot}\,\boldsymbol{A}$ (see Sect. 3.2.4) since every function $\boldsymbol{A} + \boldsymbol{u}$ with $\mathbf{rot}\ \boldsymbol{u} = \mathbf{0}$ gives the same magnetic field $\boldsymbol{B}$.

The additional ***Lorenz gauge condition***

$$\mathrm{div}\boldsymbol{A} = -\frac{1}{c^2}\frac{\partial \phi_{\mathrm{el}}}{\partial t}, \tag{4.29}$$

which reduces for stationary fields to the condition (3.12) fulfills the Maxwell Eqs. (4.25a–4.25d) as can be seen from (4.28) because

- $\mathbf{rot}\,\boldsymbol{E} = -\mathbf{rot}\,\mathbf{grad}\,\phi_{\mathrm{el}} - \mathbf{rot}\,\partial A/\partial t = -\partial B/\partial t$
  $\nabla \times \nabla\phi \equiv 0$ applies
- $\mathrm{div}\,\boldsymbol{B} = \mathrm{div}\,\mathbf{rot}\,\boldsymbol{A} \equiv 0.$

From (4.25c) we get with (4.29)

$$\begin{aligned} \mathrm{div}\,\boldsymbol{E} &= \mathrm{div}\left(-\mathbf{grad}\,\phi_{\mathrm{el}} - \frac{\partial \boldsymbol{A}}{\partial t}\right) = \frac{\varrho}{\varepsilon_0} \\ \Rightarrow \Delta\phi_{\mathrm{el}} - \frac{1}{c^2}\frac{\partial^2 \phi_{\mathrm{el}}}{\partial t^2} &\equiv -\frac{\varrho}{\varepsilon_0}. \end{aligned} \tag{4.30a}$$

This represents an extension of the Poisson equation $\Delta\phi_{el} = -\varrho/\varepsilon_0$ of electrostatics (1.16) to temporal variable fields.

For the vector potential $\boldsymbol{A}$ we obtain from (4.25b)

$$rot\ rot\,\boldsymbol{A} = \mu_0 j + \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t}. \tag{4.30b}$$

With the vector relation
$\nabla \times \nabla \times \boldsymbol{A}$ = grad div $\boldsymbol{A}$ – $\Delta\boldsymbol{A} = -(1/c^2) \cdot (\partial\phi_{\mathrm{el}}/\partial t)$ we get by inserting this into (4.30b)

$$\Delta\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} = -\mu_0 j, \tag{4.31}$$

This equation represents the extension of the Biot-Savart-law. For stationary fields it reduces to (3.13)

With the introduction of the electrodynamic potentials ϕ_{el} and $\boldsymbol{A}$ together with the Lorentz gauge it is possible to transform the coupled differential equations for $\boldsymbol{E}$ and $\boldsymbol{B}$ of first order (Maxwell equations) into uncoupled differential equations of second order for the potentials

$$\Delta\phi_{\mathrm{el}} - \frac{1}{c^2}\frac{\partial^2 \phi_{\mathrm{el}}}{\partial t^2} = -\frac{\rho}{\varepsilon_0} \tag{4.32a}$$

$$\Delta\boldsymbol{A} - \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2} = -\mu_0 j \tag{4.32b}$$

where the vector Eq. (4.32b) for $\boldsymbol{A}$ stands for three equations for the components of $\boldsymbol{A}$.

In the charge free and current free vacuum is $\rho_{\mathrm{el}} = 0$ and $\boldsymbol{j} = 0$. The equations above then reduce to

$$\Delta\phi_{\mathrm{el}} = \frac{1}{c^2}\frac{\partial^2 \phi_{\mathrm{el}}}{\partial t^2}; \tag{4.32c}$$

$$\Delta\boldsymbol{A} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{A}}{\partial t^2}. \tag{4.32d}$$

The comparison with Eq. (11.69) in Vol. 1 shows, that these equations represent waves of ϕ_{el} and $\boldsymbol{A}$ and therefore also of $\boldsymbol{E}$ and $\boldsymbol{B}$, which propagate in space as electromagnetic waves with the phase velocity $\boldsymbol{v}_{\mathrm{phase}} = c$ which equals the speed of light c (see Chap. 7).

Summary

- The temporal variation of the magnetic flux $\Phi_m = \int \boldsymbol{B} \cdot d\boldsymbol{A}$ through a coil with current I induces between the ends of the coil a voltage

$$U_{ind} = -\frac{d\Phi_m}{dt}$$

- The currents, fields and forces generated by induction are directed in such a way, that they counteract the induction process (Lenz's rule).
- The self-inductance L of an electrical network causes an induced voltage

$$U_{ind} = -L \cdot \frac{dI}{dt},$$

  which is oppositely directed to the external voltage applied to the network.
- The mutual inductance L_{12} between two conductor circuits depends on their distance and their mutual orientation.
- The spatial energy density of the magnetic field in vacuum is

$$w_{magn} = \frac{1}{2\mu_0} B^2 = \frac{1}{2} B \cdot H.$$

- The energy density of the electromagnetic field in vacuum is

$$w_{em} = \frac{1}{2} \varepsilon_0 (E^2 + c^2 B^2).$$

- The general expression of the energy density which is valid in vacuum as well as in matter reads

$$w_{em} = \frac{1}{2} (ED + BH).$$

- A temporally variable electric field $\boldsymbol{E}$ induces a magnetic field $\boldsymbol{B}$ according to

$$\mathbf{rot}\, \boldsymbol{B} = \frac{1}{c^2} \frac{\partial \boldsymbol{E}}{\partial t}.$$

- All observed phenomena of Electrodynamics can be described by the Maxwell Eqs. (4.25a–4.25d) resp. (4.26a–4.26e) and the Lorentz force (3.29b). The Maxwell equations obey the continuity equation

$$\mathrm{div}\, \boldsymbol{j} + \frac{\partial \varrho}{\partial t} = 0.$$

The Maxwell equations can be derived from general principles: The conservation of charge, the conservation of the magnetic flux and the Lorentz force on moving charges in magnetic fields. They are based on experimentally observed quantities and can be therefore tested by experiments.

Problems

4.1 A rectangular bow in the x-y-plane with width $\Delta y = b$ is placed in a homogeneous magnetic field perpendicular to the field direction (z-direction). If a rod in y-direction which touches the bow is pulled friction free with constant velocity v into the x-direction (Fig. 4.24) work has to be performed against the Lorentz force.
(a) Show that a voltage $U_{\text{ind}} = -\mathrm{d}\Phi_m/\mathrm{d}t$ is generated, that is equal to the "Hall-voltage" between the ends of the bow.
(b) Show furthermore that the mechanical power equals the electrical power $P = U \cdot I$, if the electric and mechanical resistance of the sliding rod can be neglected.
(c) The area enclosed by the bow is penetrated by an inhomogeneous field $\boldsymbol{B} = \{0, 0, B_z\}$ with $B_z = a \cdot x$. What is the time variation of the induced current $I(t)$ if the resistance of the bow $R = b \cdot L$ is proportional to the total length L of the bow.

4.2 Calculate the self-induction per meter for a cable consisting of two concentric cylindrical conducting tubes with radii R_1 and R_2 for the back and forth current. What is the magnetic energy density in the space between the two tubes with the current I?
Numerical example: $R_1 = 1$ mm; $R_2 = 5$ mm; $I = 10$ A.

4.3 Two concentric circular rings which both lie in the same plane have the radii R_1 and R_2.
(a) What is the mutual induction?
(b) What is the induction flux Φ_{m}, when one of the rings carries the current I? Show, that Φ_{m} does not depend on the choice which of the rings carries the current I.

4.4 A two-conductor line consists of two thin stripes with width $b = 10$ cm, and the thickness $d = 0.1$ cm at a distance of 0.2 cm. A current I flows through each of the two lines into opposite direction. Calculate the induction L and the capacity per m length, when between the two stripes an isolation material with $\varepsilon = 5$ is located. Does the product $L \cdot C$ depend on the dimensions of the double conductor line?

4.5 Show, that for the Waltenhofen-pendulum in Fig. 4.12 the damping torque
(a) $D_0 \propto \mathrm{d}\varphi/\mathrm{d}t$, where φ is the angle of the pendulum bar with the vertical direction
(b) $D_0 \propto I^2$ where I is the current through the field coils.

4.6 A switch opens the connection of a coil ($L = 0.2$ H, $R_{\text{L}} = 100\ \Omega$) with the voltage supply ($U_0 = 20$ V). Calculate the current $I(t)$.

4.7 Show with Gauss's Law, that the temporal change $\mathrm{d}Q/\mathrm{d}t$ of the charge $Q = \int \varrho \cdot dV$ in the volume V and the current $I = \int j \cdot \mathrm{d}\boldsymbol{S}$ through the surface S enclosing the volume V obeys the continuity equation

$$\dot{\varrho} + \mathrm{div}\, j = 0$$

4.8 A train runs with the speed $v = 200$ km/h over a straight railroad where the rails have a distance of 1.5 m. Which voltage is generated between the two rails due to the earth magnetic field $B = 4 \cdot 10^{-5}$ T when $\boldsymbol{B}$ is inclined against the vertical direction by 65°?

4.9 A coil with N windings encloses a straight wire, carrying an ac current $I = I_0 \cdot \sin \omega t$. What is the voltage induced across the ends of the coil
(a) if the N windings form concentric circles around the wire
(b) if the coil windings form a torus and its central line a circle with radius R_2 around the wire
(c) if a rectangular flat coil with N windings and a side length a in the radial direction and the side length b parallel to the wire is placed at a distance d resp. $d + a$ from the wire?

4.10 An electromagnet is fed with a current $I = 1$ A, which flows through 10^3 windings of the field coil with a cross section of 100 cm^2, a length $l = 0.4$ m and an electrical resistance of $R = 5\ \Omega$. The magnetic field in the iron rod is $B = 1$T. What is the induction voltage across the ends of the coil, if the current is shut off within the time $\Delta t = 1$ ms?

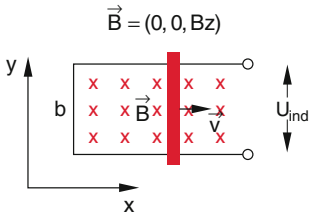


Fig. 4.24 To problem 4.1

References

1. J.V. Parker: Electromagnetic Projectile Acceleration. J. Appl. Phys. 53, 6711 (1982)
2. https://en.wikipedia.org/wiki/Railgun
3. https://en.wikipedia.org/wiki/Magnetic_levitation
4. John Waymouth, Electric Discharge Lamps. Cambridge, MA: The M.I.T. Press 1971
5. J.D. Jackson: Classical Electrodynamics 3rd ed. (Wiley 1998)
6. W. Greiner: Classical Electrodynamics (Springer, New York 1998)
7. D.J. Griffith: Introduction to Electrodynamics (Cambridge Univ. Press 2017)

Electrotechnical Applications

5

Fundamental research about electric and magnetic fields and their temporal changes had already in the 19th century opened the way to many technical applications, which had essentially contributed to the "technical revolution" and which had changed public life in many ways. Examples are the generation and transportation of electric energy and its applications in industry, traffic and private households.

In this chapter we will discuss only some of the most important applications which are still in use today.

5.1 Electric Generators and Motors

Faraday's Induction law represents the basis of electric generators.

The most simple model of an ac-generator is a rectangular conduction loop with cross section area A and N windings which rotates in a homogeneous magnetic ac-field $B = B_0 \cos \omega t$ with the angular frequency ω (Fig. 4.3). It generates the induced voltage

$$U = B_0 \cdot N \cdot A \cdot \omega \cdot \sin \omega t,$$

which is transferred through two sliding contacts K_1 and K_2 to fixed output contacts (Fig. 5.1).

A generator transforms mechanical energy (which is needed for turning the loop) into electric energy. On the other hand the generator can be also used as a motor: When an external ac voltage is supplied to the output contacts of the generator, the loop turns with the frequency of the external ac voltage and the system converts electric into mechanical work. The generator has changed to a motor.

If a dc-current is sent through the loop, it can perform only half a turn. If now the magnetic field is always commutated at the right times, the loop can rotate continuously. This commutation is realized by a slotted sliding contact called the **commutator** (Fig. 5.2a). For the case of a single coil it consist of two halves which are isolated against each other but connected to the two ends of the coil. The commutator allows the use of the generator as direct current (dc) generator or motor.

When the loop with the commutator is turned by hand, the output provides a pulsed dc-voltage (Fig. 5.2b). This pulsating output can be smoothened by using N loops which are tilted against each other by the angle $\alpha = \pi/N$. The commutator now consists of N segments with N output contacts. The end of each loop is connected to the beginning of the next loop and to the corresponding segment of the commutator. For illustration a generator with two loops is shown in Fig. 5.3a. The output voltage of the two loops behind the commutator, which are shifted against each other by half a period, and the sum of the two voltages are depicted in Fig. 5.3c. The electric circuitry is illustrated in Fig. 5.3b.

When the generator is used as a motor the magnetic force acting onto the coils which carry the current I can be greatly enhanced when a cylindrical iron core is inserted between the two circularly carved poles of the magnet (Fig. 5.4). The magnetic field in the gap between the iron core and the pole

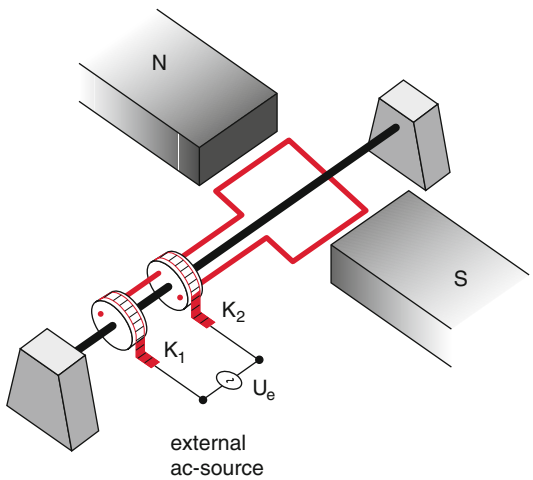


Fig. 5.1 Simple model of an ac-generator

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_5

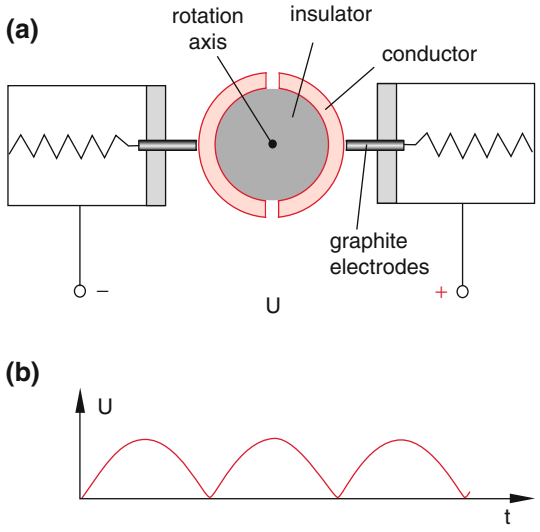


Fig. 5.2 **a**) dc-generator resp. motor with two-part commutator and sliding contacts, **b**) pulsating dc-voltage with only one rotating coil

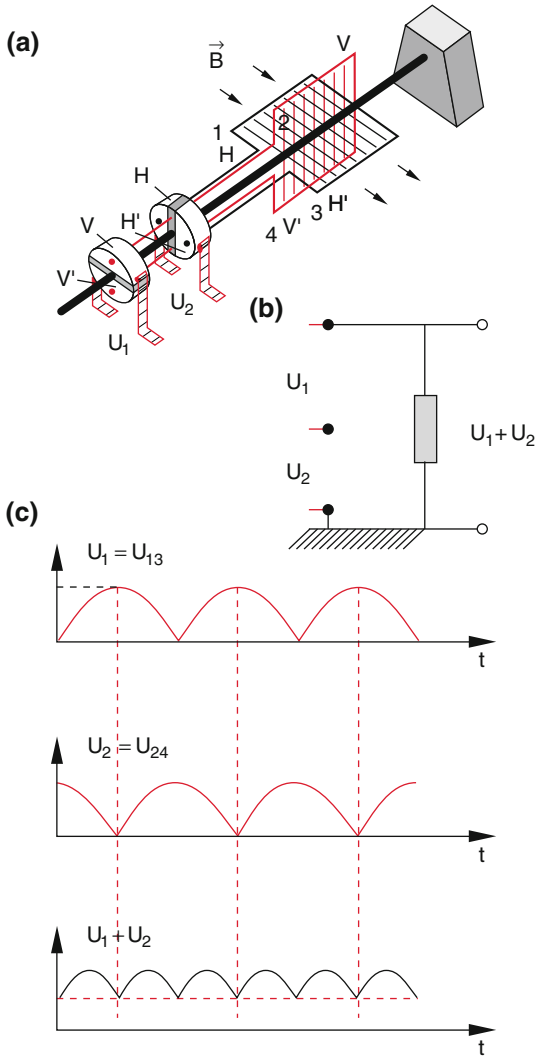


Fig. 5.3 **a**) Generator with two coils. *H* and *H′* are the two sides of the horizontal coil, *V* and *V′* those of the vertical coil

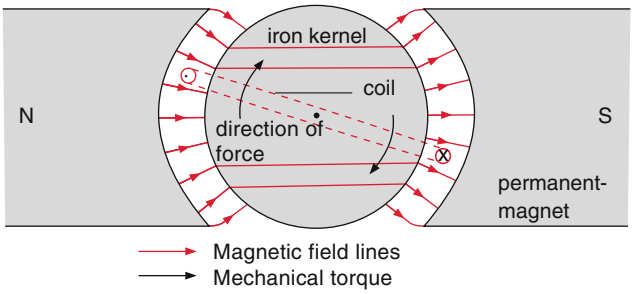


Fig. 5.4 Amplification of the force acting on the current loop by an iron kernel generating a radial magnetic field

shoes is radial. Therefore the force is approximately constant for the whole time when the coil rotates in the gap.

The three most important parts of a generator (or motor) are

- the fixed field magnet (stator)
- the rotating coils (rotor)
- the commutator or collector with the sliding contacts, which are realized by carbon rods pressed by springs onto the collector (Fig. 5.2a).

The optimization of the rotor was achieved by the invention of the drum armature. Its principle is illustrated in Fig. 5.5. Instead of the coils a cylinder of magnetic material is used where the coils are winded in grooves milled into the iron cylinder (Fig. 5.5b). This greatly enhances the magnetic flux through the coils. For *N* segments on the drum collector

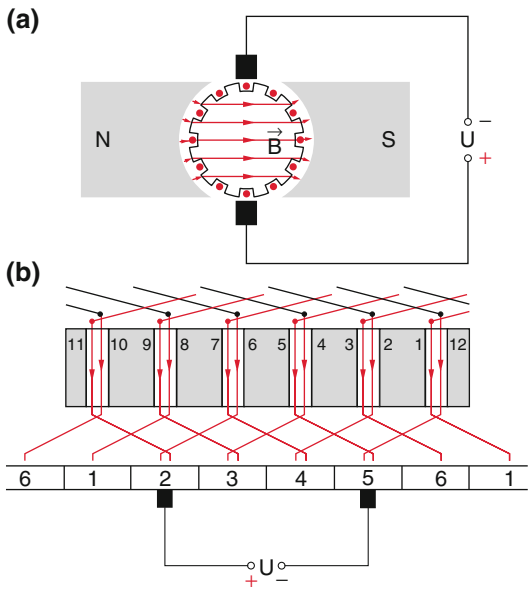


Fig. 5.5 **a**) Drum armature with a magnetizable iron core. Six coils (red points) are embedded into grooves of the core. **b**) Description of the connections between the coils presented in a plane. The numbers give the corresponding grooves. The carbon rods shown in (**a**) form sliding contacts on the collector behind the drum armature

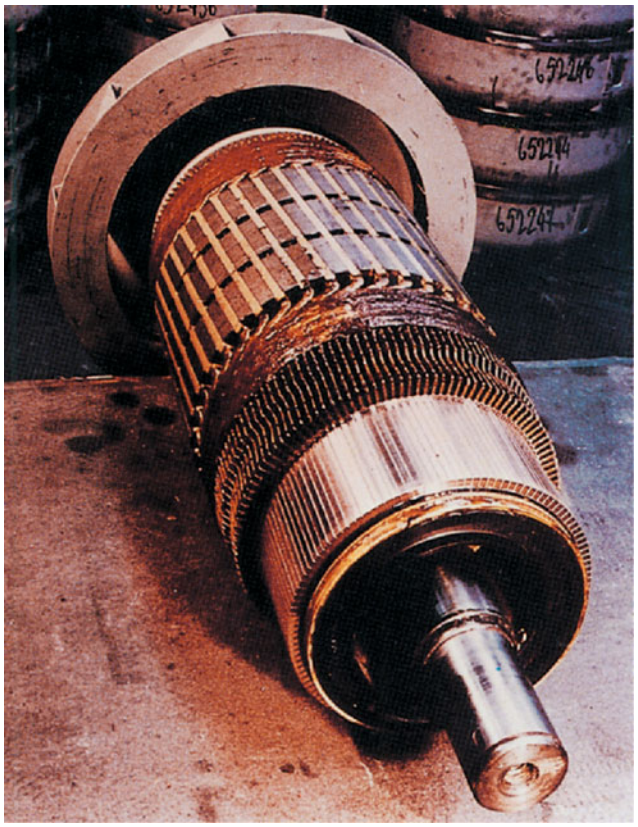


Fig. 5.6 Rotating part of a dc-motor with commutator, armature windings and fan propeller, (with kind permission of Siemens AG)

the voltages of the different coils have to be added with the correct phase. This can be reached, if the end of a coil is connected with the beginning of the next coil. For the coil position, shown in Fig. 5.5. The voltage between the segments 2 and 5 just gives the voltage between the ends of that coil with its area perpendicular to the magnetic field of the stator. Figure 5.6 shows a commercial dc-machine, which can be used as generator as well as motor.

Since the induced voltage is proportional to the magnetic field B the stator field should be as high as possible. This can be best achieved with electromagnets instead of permanent magnets. In order to save an external power supply, all electrical machines produce their own field current. They use the fact, that electromagnets have, even without current, a residual magnetic field due to the remnant magnetization of iron (see Fig. 3.45, remanence). This residual field is sufficient to induce a voltage when the coils on the rotor turn. The resultant current is used to enhance the magnetic field of the stator. This **dynamo-electric principle** was discovered in 1866 by *Werner von Siemens*. It allows the construction of large generators which do not need any external power supplies. The enhancement of the magnetic field is limited by saturation effects and by Ohmic losses in the windings. Generators based on the dynamo-electric principle are called **dynamos**.

Note, that a higher electrical output power of generators demands a higher mechanical input power for driving the generator. This is illustrated by Fig. 5.12 where a high power gas turbine is used for driving an electric generator.

The energetic efficiency of a generator is defined as the ratio of electric output power to mechanical input power. It is always smaller than one because of the unavoidable losses by the ohmic resistance and by mechanical friction.

5.1.1 DC-Machines

Depending on the specific applications three different circuits of dc-machines are used:

5.1.1.1 Series Wound Motor

In the series wound motor (Fig. 5.7) the total current produced in the rotor coils is sent after rectification by the commutator through the stator coils and the load resistor R_a. This means that rotor, field coils in the stator and load resistor are connected in series. The magnetic field current is equal to the load current.

With increasing current I the magnetic field increases and therefore also the induced voltage. Because of saturation in the magnetic iron core the line $U = f(I)$ is curved (Fig. 5.7c). Stationary operation is reached at the crossing point of the straight line $U = (R_i + R_a)\ I$ with the curved line $U = f(I)$. With the total internal resistance $R_i = R_F + R_R$ as the sum of the resistance of the field coils and that of the rotor the terminal voltage is

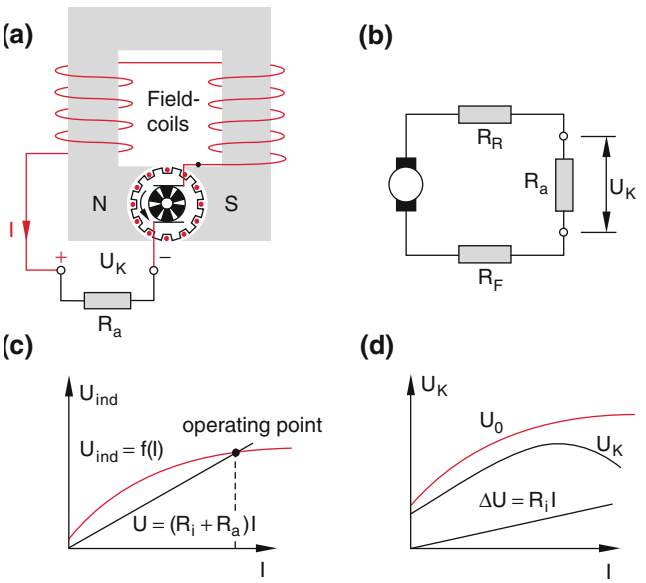


Fig. 5.7 Motor with series winding. **a**) Schematic representation, **b**) equivalent circuit R_R is the resistance of the rotor coils, R_F that of the field coils, R_a the load resistor, **c**) excitation curve with operating point, **d**) current-voltage characteristic

$$U_K = U_0 - I \cdot R_i, \tag{5.1}$$

where U_0 is the voltage for $R_i = 0$. If the resistance R_R of the rotor is very small compared with the resistance R_F of the field coils U_0 is nearly equal to the induced voltage U_{ind} and therefore proportional to the magnetic field B and the field current I. In Fig. 5.7d the terminal voltage U_K and the voltage U_0 are plotted as a function of the current I. Because of saturation of the magnetic field for high currents U_0 approaches a constant value and U_K decreases according to (5.1).

With a load resistor R_a the total output power of the series wound generator (Hauptschluß-Machine).

$$P = U_0 \cdot I = I^2 \cdot (R_i + R_a),$$

where $R_i = R_R + R_F$ is the total inner resistance (rotor and field coils). The fraction $P_i = I^2 R_i$ is consumed inside the machine and only the part $P_a = I^2 R_a$ is supplied to the external load. The electric efficiency of the hauptschluss-machine is

$$\eta = \frac{P_a}{P} = \frac{R_a}{(R_i + R_a)}. \tag{5.2}$$

In order to realize a maximum efficiency the internal resistance R_i should be as small as possible. This means one has to use thick wires for the coils.

The advantage of the hauptschluss-machine is its adaption to the power consumed by external consumer. If a larger power is consumed the current I increases and therefore the magnetic field and the available power of the machine. Its disadvantage is that the supplied voltage is not constant.

5.1.1.2 The Shunt-Motor

In the shunt motor (parallel circuit Fig. 5.8) the external load circuit and the magnet coils are connected in parallel. Even without load the current for the magnetic field I_F remains constant. The current, taken from the commutator

$$\begin{aligned} I &= I_F + I_a = \frac{U_{ind}}{R_F} + \frac{U_{ind}}{R_a} \\ &\Rightarrow I_F/I_a = R_a/R_F \end{aligned} \tag{5.3a}$$

is split into the field current I_F and the consumer current I_a. With $I = I_F + I_a$ we obtain

$$I_a = I \cdot \frac{R_F}{R_a + R_F}; \quad I_F = I \cdot \frac{R_a}{R_a + R_F} \tag{5.3b}$$

The electric power delivered to the consumer is $P_a = I_a^2 \cdot R_a$, the power consumed in the field magnet $P_F = I_F^2 \cdot R_F$ and in the rotor $P_{R\vartheta} = I^2 \cdot R_R$. The electric efficiency is then

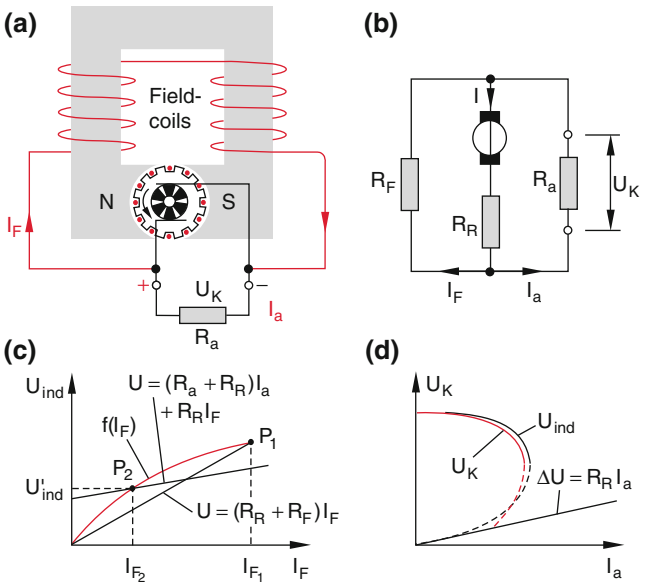


Fig. 5.8 **a**) Motor with parallel winding, **b**) schematic representation, **c**) excitation curve with load-dependent operation point, **d**) current-voltage characteristic

$$\begin{aligned} \eta &= \frac{P_a}{P_a + P_F + P_R} \\ &= \frac{I_a^2 R_a}{I_a^2 R_a + I_F^2 R_F + I^2 R_R}. \end{aligned} \tag{5.4}$$

With (5.3a, 5.3b) this gives

$$\eta = \frac{1}{1 + \frac{R_R}{R_a} + \frac{R_a + 2R_R}{R_F} + \frac{R_a R_R}{R_F^2}}, \tag{5.5}$$

This shows that for optimization of η the resistance of the field coils should be as large as possible, contrary to the hauptschluß- machine where it should be as small as possible.

The current-voltage characteristic of a shunt-machine is shown in Fig. 5.8d. Without load ($I_a = 0$) is $I = I_F$. The total current delivered by the rotor flows through the field coils. This causes a maximum of the induced voltage $U_{ind} = U_1$. It adjusts to the intersection P_1 between the straight line.

$$U = (R_R + R_F) \cdot I_{F_1}$$

and the curve $U = f(I_F)$ which is determined by the magnetization of the magnet.

If now a consumer is connected parallel to the inner circuit the current $I = I_F + I_a$ increases. This causes a decrease of the terminal voltage to the lower value

$$U_K = U_{ind} - R_R(I_F + I_a). \tag{5.6}$$

With decreasing voltage U_K also the field current $I_F = U_K/R_F$ and the magnetic field B decrease. This reduces the induced voltage to the value

$$U'_{ind} = f(I_{F_2}) = U_2 = (R_a + R_R)I_a + R_R I_{F_2} = (R_a + R_R)I - R_a I_{F_2} \quad (5.7)$$

which corresponds to the point P_2 in Fig. 5.8c.

With increasing load current I_a the straight line U_2 is shifted. Above a critical current I_a there is no longer a crossing point, which means that in this range a stable operation of the machine is no longer possible.

The shunt machine is generally operated in the upper part of the $U(I_2)$ characteristic in Fig. 5.8d. If the output terminal are short circuited $(R_a = 0)$ the voltage U and the slope dU/dI_2 become zero. Therefor a short circuit does not harm the machine.

The advantage of the shunt machine is a good stability of the output voltage in the upper part of the current-voltage characteristic. Its disadvantage is the small resistance against changes of the load. If the load current becomes too large, the machine may stand still already at lower load powers than for the hauptschluss engine.

5.1.1.3 The Compound Motor

The advantages of the series wound motor and the hauptschluss and the shunt machine can be combined by introducing two separate coils for the magnet: One with thick wires (small resistance) which is arranged in series with the load circuit and one with large resistance R_F which is connected as parallel circuit to the coil with low resistance (Fig. 5.9). This gives a better stability of the voltage $U_2(I_a)$ and furthermore a better adaption to strongly changing load conditions.

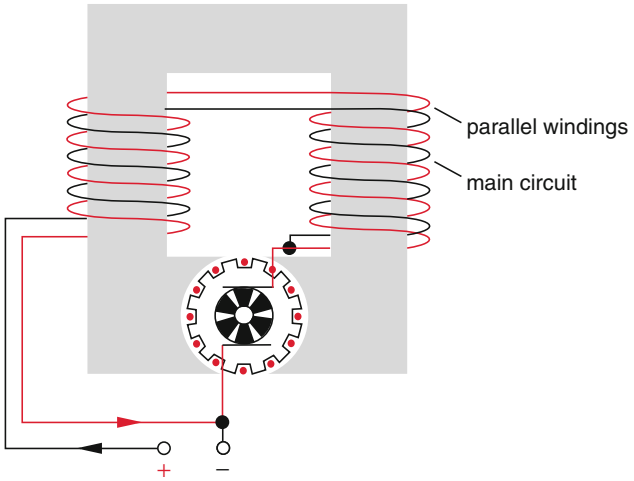


Fig. 5.9 Compound motor

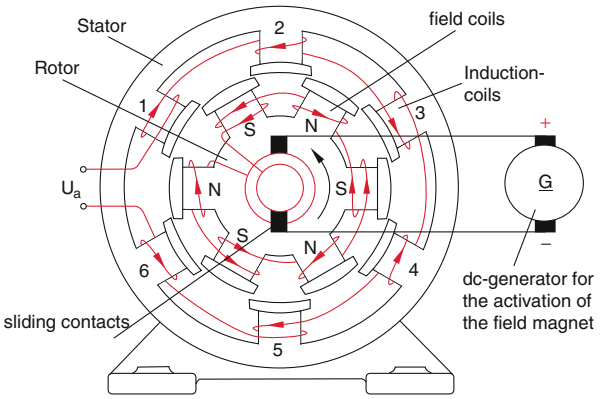


Fig. 5.10 Ac-generator with rotating field magnet and fixed induction coils and external dc-generator for supplying the current for the field magnets

5.1.2 AC-Generators

AC-machines do not need the commutator. The simple model of Fig. 4.5 is, however, modified for obtaining a higher efficiency.

Nowadays most of the generators are internal pole machines, where the magnetic field coils rotate and the induction coils are fixed. The advantage of this design is that no sliding contacts are necessary which always represents a problem for the transfer of large currents to the consumer. In Fig. 5.10 a six-pole ac generator is shown as example. The rotating field magnet is an electromagnet with three north- and three south-poles. Fixed at the casing are six induction coils with iron kernel and alternate reversed coiling direction and connected in series. The output voltage at the terminals is the sum of the three induction voltages.

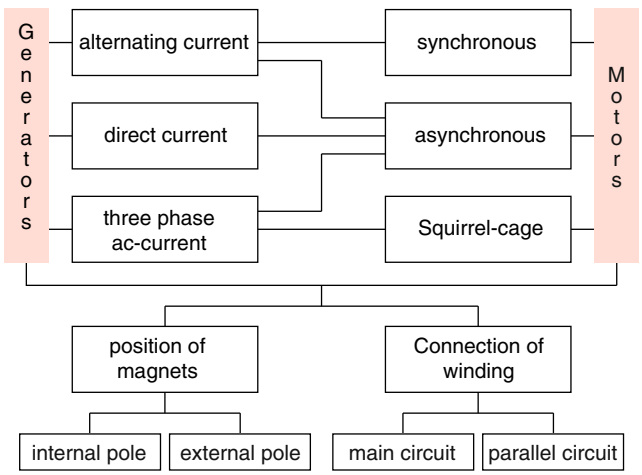


Fig. 5.11 Survey about the different designs of generators and motors

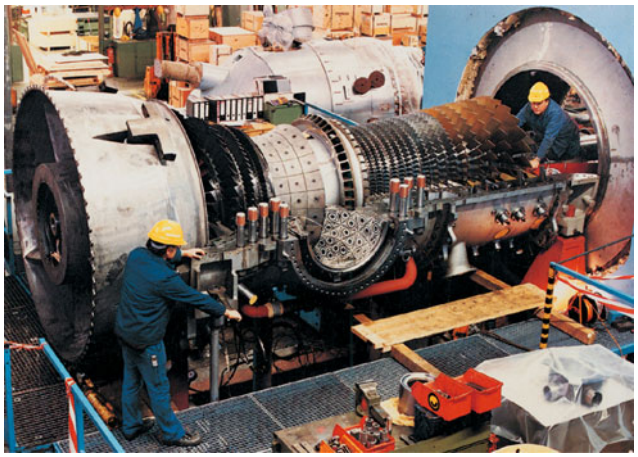


Fig. 5.12 New high temperature gas turbine for driving a high power generator (with kind permission of Siemens AG)

The rotor magnets receive the much smaller current over sliding contacts. The voltage can be either supplied as shunt circuit from the terminals and is then rectified, or an extra dc-generator is installed which delivers the voltage to the coils.

In order to generate an ac current with a frequency of 50 Hz the rotation frequency of the three induction coils has to be $f = 50 \cdot 60/3 = 1000$ turns per minute Fig. 5.11 gives a compilation of the different types of generators and motors and Fig. 5.12 illustrates the size of a gas turbine with 450 MW output for driving a 400 MW generator.

There are two different kinds of ac-generators: **Synchronous generators** where the rotation frequency $f = f_a/n$ for n poles is synchronized with the frequency f_a of the external ac-voltage and **asynchronous machines** which are generally operated as three phase ac- generators (see Sect. 5.3) and which run at a rotation frequency $f = f_a/n$ that is smaller than that of the external voltage. Due to their robust operation and their simpler setup nowadays mainly asynchronous machines are used. In Fig. 5.13 a big three-phase generator with an output power of 100 MW is shown during its mounting. In the Diagram 5.11 the different technical designs for motors and generators are compiled [1].

5.2 Alternating Current (AC)

The alternating voltage

$$U(t) = U_0 \cdot \cos\ \omega t,$$

across a resistor R causes an alternating current

$$I(t) = I_0 \cdot \cos\ \omega t \quad \text{with} \quad I_0 = U_0/R$$

The time interval $T = 2\pi/\omega$ between two maxima is the period of the ac current (Fig. 5.14). In the integrated network

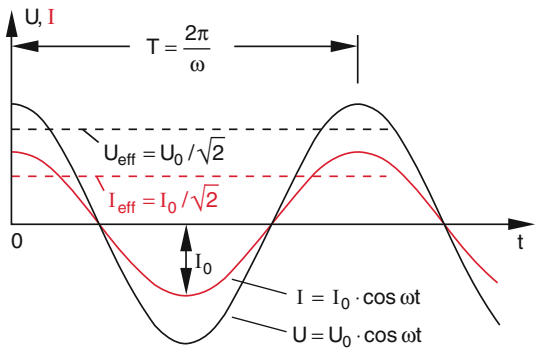


Fig. 5.14 The characteristic features of the ac-voltage

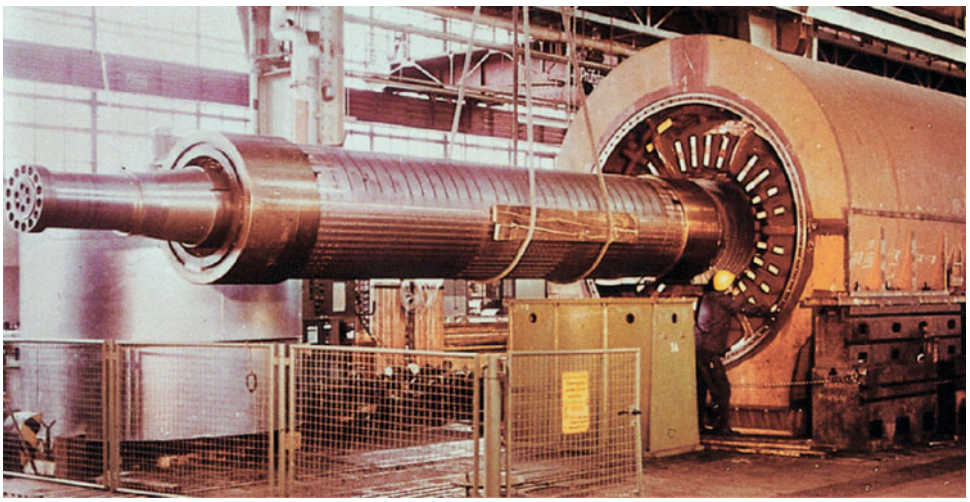


Fig. 5.13 Installation of the rotor into a three-phase ac generator with 3000 turns per minute and an output power of 100 MW. (with kind permission of Siemens AG)

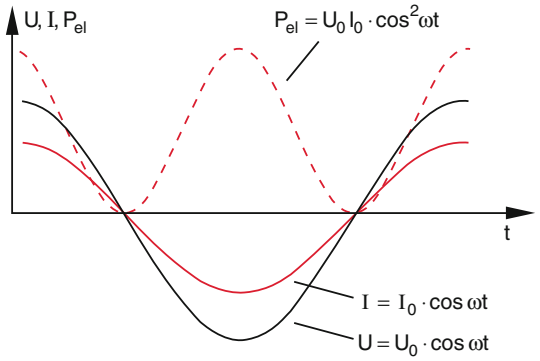


Fig. 5.15 Power curve of dc-current for in-phase of current and voltage

of Europe is $\omega = 2\pi \cdot 50\,\text{Hz} \Rightarrow T = 20\,\text{ms}$. The electric power of this current and voltage

$$P_{\text{el}} = U \cdot I = U_0 I_0 \cos^2 \omega t \tag{5.8a}$$

is also a periodic function of time (Fig. 5.15). Its time average is

$$\begin{aligned} \bar{P}_{\text{el}} &= \frac{1}{T}\int_0^T U_0 I_0 \cos^2 \omega t \, \mathrm{d}t \quad \text{with} \quad T = 2\pi/\omega \\ &= \frac{1}{2} U_0 I_0. \end{aligned} \tag{5.8b}$$

A dc-current $I = I_0/\sqrt{2}$ caused by a dc-voltage $U = U_0/\sqrt{2}$ has the same time averaged power as the ac-voltage and current with the amplitudes (maximum values) U_0 and I_0. The expressions

$$U_{\text{eff}} = \frac{U_0}{\sqrt{2}} \quad \text{and} \quad I_{\text{eff}} = \frac{I_0}{\sqrt{2}} \tag{5.9a}$$

are therefore called the **effective or actual values** (root mean square values) of voltage and current (Fig. 5.16).

Example

For the European uniphase network we get at our electrical outlet the effective voltage $U_{\text{eff}} = 230\,\text{V} \Rightarrow U_0 = 230 \cdot \sqrt{2}\,\text{V} = 325\,\text{V}$. With $f = \omega/2\pi = 50\,\text{Hz} \Rightarrow \omega = 300\,\text{s}^{-1}$. We can therefore write:

$$U(t) = 325 \cdot \cos\left(2\pi \cdot 50 \cdot t/\text{s}\right)\text{V}.$$

If the network includes inductances L or capacitors C, voltage and current are generally not in phase (see Sect. 5.4) but show a phase difference φ:

$$U = U_0 \cdot \cos \omega t, \quad I = I_0 \cdot \cos\left(\omega t + \varphi\right). \tag{5.9b}$$

The average power is then

$$\begin{aligned} \bar{P}_{\text{el}} &= \frac{U_0 I_0}{T}\int_0^T \cos \omega t \cdot \cos(\omega t + \varphi)\, \mathrm{d}t \\ &= \frac{U_0 I_0}{2} \cdot \cos \varphi. \end{aligned} \tag{5.10}$$

For $\varphi = 90°$ the average power becomes zero (Fig. 5.17).

Example

An inductance L in the network with an ohmic resistance $R = 0$ does not consume, on the time average, any energy. The energy needed for building up the magnetic field during half a period of the ac power, is released again in the next half, when the magnetic field decays.

Corresponding considerations are valid for a capacitor in the network.

The power taken and released again by inductances and capacitors is therefore called **idle power** or **wattless power**, while the actual in ohmic resistors consumed power is the **real or wattful power** (Figs. 5.17 and 5.18).

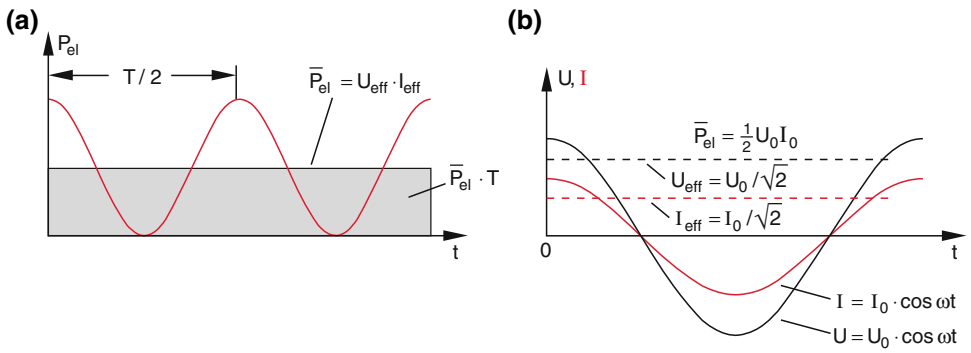


Fig. 5.16 **a**) Average power of ac-current. **b**) Effective values of current and voltage

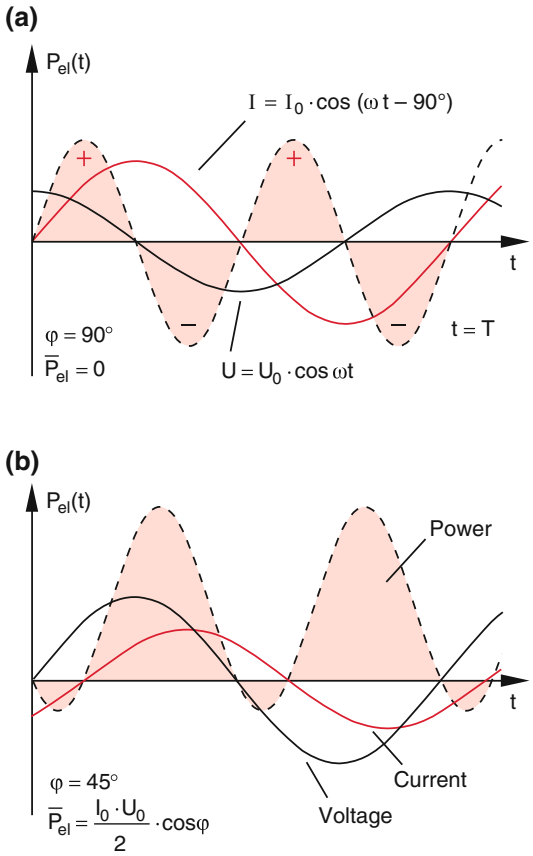


Fig. 5.17 Temporal dependence of electric power $P = I \cdot U$ for different values of the phase shift φ between current I and voltage U. The effective power is the difference between the red shaded areas above and below the t-axis. In **a** is the effective power $\bar{P}_{el} = 0$ in **b** is $P_{el} \neq o$

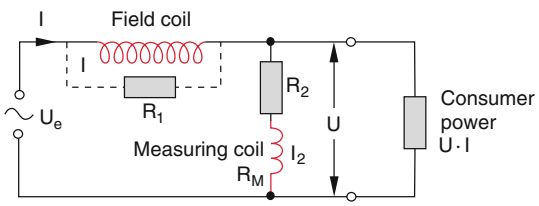


Fig. 5.18 Circuit for measuring the effective power

For measurements of the wattful power instruments are used where the readout is proportional to the effective power $\bar{P}_{el}$. One example is the circuit of a modified moving coil meter shown in Fig. 2.28. The permanent magnet in Fig. 2.28 is replaced by a fixed field coil. Through this coil the measured load current I flows. The field coil provides the magnetic field for a moving coil galvanometer while the measuring coil represents the moving coil of the instrument. The voltage $U = I_2 \cdot (R_2 + R_m)$ is measured through the current I_2 that flows through the measuring coil with resistance R_m and a large series resistor R_2 limiting I_2 to values $I_2 << I$. The magnetic moment of the measuring coil is proportional to the voltage U and the magnetic field of the field coil is proportional to the current I: Therefore the acting torque on the measuring coil is proportional to the product $U \cdot I$. The mechanical inertia of the pointer connected to the measuring coil and indicating on a scale the measured values prevents that it follows the fast oscillations of the ac power and therefore indicates the average power. The measuring range of the device can be enlarged by parallel resistors or resistors in series.

5.3 Multiphase and Rotary Currents

Replacing the coil in Fig. 5.1 by N coils which are twisted against each other by the angle $2\pi/N$, the voltages measured between the ends of each coil

$$U_n^{\text{ind}} = U_0 \cos\left(\omega t - \frac{n-1}{N} \cdot 2\pi\right) \tag{5.11}$$

are shifted by the angle $\Delta\varphi = 2\pi/N$. These voltages can be measured if one end of all coils is connected to the same sliding contact and the other ends to N different contacts. The device has then $N + 1$ contacts.

For technical applications the three-phase current with $N = 3$ has found the widest distribution because it allows with reasonable expenditure the transport of electric energy with a given electrical power.

A general method for the generation of a three-phase current uses a magnet that rotates about a central axis in Fig. 5.19. It induces in three fixed coils which are displaced by 120° three ac-currents that are also shifted against each other by 120°. Connecting one end of each coil to a resistor R_i(i = 1, 2, 3), and the other ends of all coils to a common return circuit (*star connection*) (Fig. 5.20), the currents $I_i = U_{\text{ind}}/R$ in the three lines i = 1, 2, 3 are

$$\begin{aligned} I_1 &= I_0 \cos \omega t, \\ I_2 &= I_0 \cos(\omega t - 120^\circ), \\ I_3 &= I_0 \cos(\omega t - 240^\circ). \end{aligned} \tag{5.12}$$

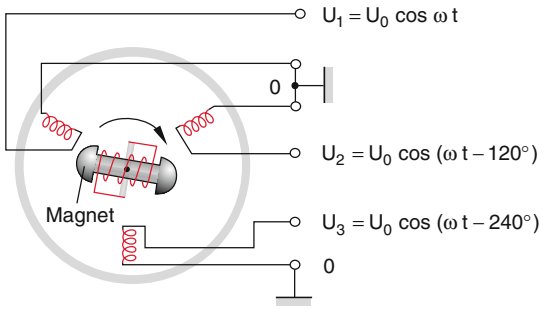


Fig. 5.19 Generation of three ac-voltages with a common reference pole 0 shifted against each other by 120° by a rotating magnet

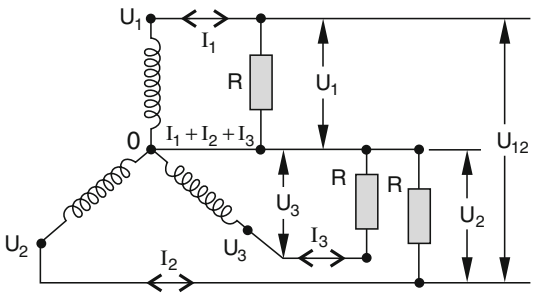


Fig. 5.20 Star connection for the three-phase ac current

Using the addition theorem for goniometric functions it follows for the sum

$$\Sigma I_i = 0 \tag{5.12a}$$

This implies that the current through the common return circuit is zero (Fig. 5.21). It is therefore called *zero conductor*. Equation (5.12a) is only valid, if each of the three lines contains the same resistor. Different resistors in the three lines change the amplitude I_0. Furthermore phase shifting elements, such as inductances or capacitors shift the relative phase between the three lines and then their currents no longer add up to zero!

When each induction coil in Fig. 5.20 delivers the same voltage amplitude U_0 the voltage between the outlets 1 and 2 is

$$\begin{aligned}\Delta U_{1,2} &= U_1 - U_2 \\ &= U_0[\cos \omega t - \cos(\omega t - 120^\circ)] \\ &= -U_0 \cdot \sqrt{3} \sin(\omega t - 60^\circ) \\ &= +U_0 \cdot \sqrt{3} \cos(\omega t + 30^\circ).\end{aligned} \tag{5.13}$$

The phase difference between outlets 1 and 2 is $\Delta\varphi_{1,2} = -30^\circ$ and the amplitude is $U_{1,2} = U_0 \cdot \sqrt{3}$. Instead of -30° the phase shift between 2 and 3 is $\Delta\varphi_{2,3} = -90^\circ$ and between 3 and 1 $\Delta\varphi_{3,1} = -150^\circ$.

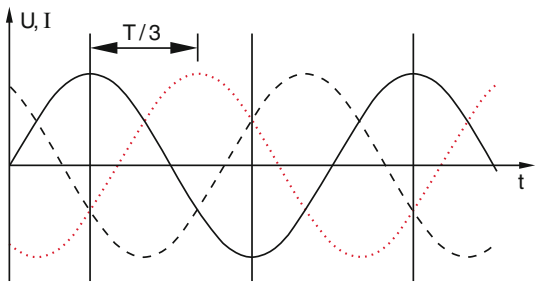


Fig. 5.21 The sum of three one-phase ac currents shifted by 120° against each other is zero indicated by the t-axis $U = 0$, $I = 0$

This shows that the voltage between two phases of the three-phase current is larger by the factor $\sqrt{3}$ than the voltage between one phase and the zero return line.

Example

$$\begin{aligned}U_1^{\text{eff}} = U_2^{\text{eff}} &= 230\,\text{V} \Rightarrow U_0 = \sqrt{2} \cdot U^{\text{eff}} \\ &= 325\,\text{V} \Rightarrow \Delta U_0 = \sqrt{3} \cdot U_0 = 563\,\text{V}.\end{aligned}$$

The maximum amplitude $\Delta U(t)$ of the ac voltage between two phases of the three phase current is 563 V, the effective value is $\Delta U_{\text{eff}} = 398\,\text{V}$.

Besides the Y-connection discussed before often the delta connection in Fig. 5.22 is used. It exploits the fact that the sum of all three voltages of a three phase current is zero.

$$U_{\text{tot}} = \sum_{n=0}^{2} U_0 \cos\left(\omega t - n\frac{2}{3}\pi\right) = 0. \tag{5.14}$$

For the delta connection the voltages between the points 1, 2, 3 are always the voltages of one phase. The advantage compared with the one-phase current is the smaller load per phase for a given output power, if the different consumers are equally distributed among the three phases. However, the current through each of the three lines is always the sum of two load currents (Fig. 5.22), which are generally phase shifted against each other. For example the current I from the outlet 1 in Fig. 5.22 is

$$I = I_1 + I_2 = U_{13}/R_1 + U_{12}/R_2.$$

For the Δ-connection is always $\sum U_i = 0$ independent of the load resistors R_i. It is, however, no longer $\sum I_i = 0$.

When the voltages

$$U_n = \sum U_0 \cos\left(\omega t - n\frac{2}{3}\pi\right),$$

obtained from the three output terminals in Fig. 5.20 are applied to three magnetic coils with axis twisted by 120° against each other, the superposition of the three fields gives a magnetic field which rotates with the frequency ω about

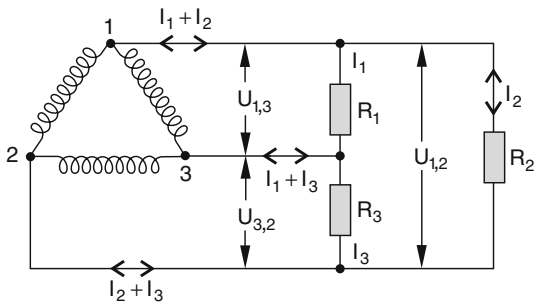


Fig. 5.22 Delta connection for the three phase ac current

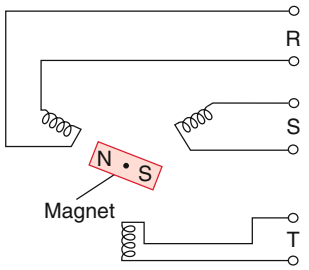


Fig. 5.23 Demonstration of the magnetic rotating field generated by the three phase ac current

the symmetry axis of Fig. 5.23 perpendicular to the drawing plane. This can be demonstrated by a compass needle supported by a pin in the symmetry axis, if the frequency ω is chosen so low that our eye can follow the rotation of the needle.

The rotation of the magnetic field can be explained by a simple vector model (Fig. 5.24). The three magnetic fields B_n of the three coils point into the direction of the coil axis (see Sect. 3.2.6.4 and Fig. 3.6). They all lie in the same plane (the drawing plane of Fig. 5.24) turned against each other by the angle 120°. Assume that at time $t = 0$ the current through coil 1 has reached its maximum and the magnetic field points radially towards the center. Since the currents in the coils 2 and 3 are phase shifted by $\pm 120°$ the fields B_2 and B_3 are weaker by the factor $\cos 120° = -1/2$, resp $\cos 240° = -1/2$. They are both directed radially outwards. The superposition of the three fields gives a field in the direction of the coil axis 1. After $1/3$ of a period, i.e. at $t = 2\pi/(3\omega)$ the total magnetic field has turned by 120° and points into the direction of the axis of coil 2 towards the center.

Because of the rotation of the magnetic field the three phase current is also called **rotary current**.

The rotating magnetic field is used for the construction of rotatary field motors [4]. Their principle was already illustrated by the rotating magnetic compass needle. The technical realization is shown in Fig. 5.25, Instead of the needle a rotary iron ring is used, which is wrapped around with coils (squirrel cage rotor).

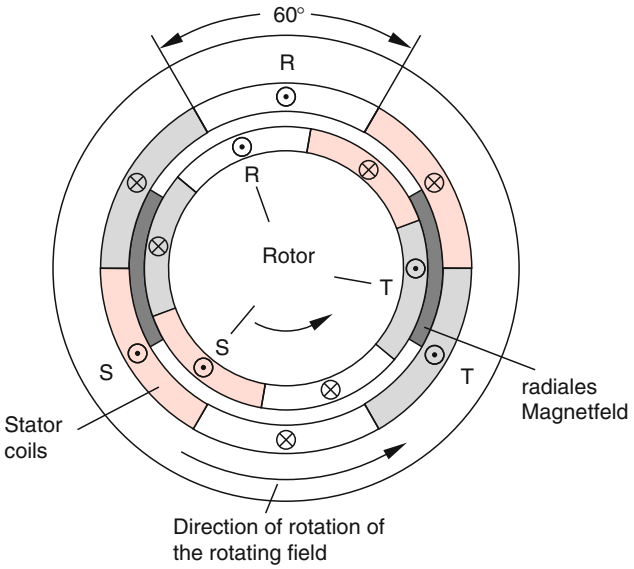


Fig. 5.25 Model of the three phase motor

5.4 AC-Current Circuits with Complex Resistors; Phasor Diagrams

The phase shifts between currents and voltages caused by inductances and capacitors in electric circuits, can be best illustrated by using a complex notation [5]. How this complex notation is translated into real circuits will be exemplified in the next section.

5.4.1 AC-Circuit with Inductance

The external voltage $U_e = U_0 \cos \omega t$ in Fig. 5.26 must be opposite to the induced voltage $U_{ind} = -L \cdot dI/dt$ and with

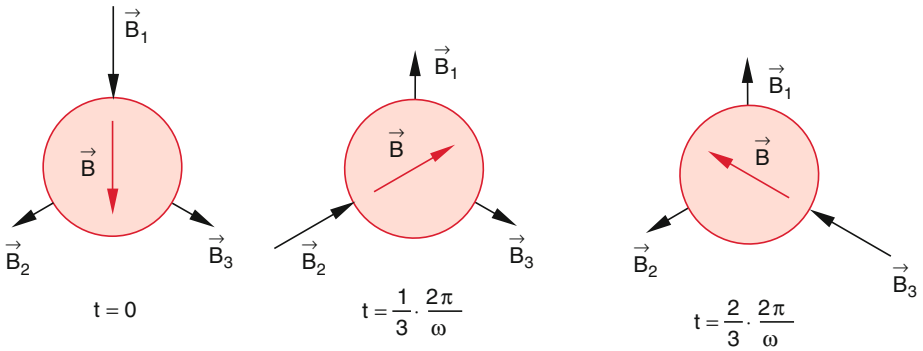


Fig. 5.24 Vector addition of the magnetic fields in the three coils of the three phase current

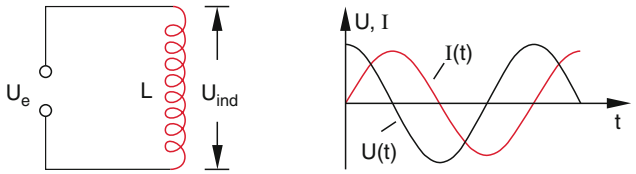


Fig. 5.26 AC circuit with inductance

equal amount, because the total voltage in a closed circuit must be zero. We will at first neglect any ohmic resistor.

$$U_e + U_{ind} = 0 \Rightarrow U_0 \cos \omega t = L \cdot \frac{dI}{dt}, \tag{5.15}$$

$$\Rightarrow I = \frac{U_0}{L} \int \cos \omega t \, dt = \frac{U_0}{\omega L} \sin \omega t = I_0 \sin \omega t \quad \text{with} \quad I_0 = \frac{U_0}{\omega L}. \tag{5.16}$$

Current and voltage are no longer in phase. The ac-current is delayed by 90° against the voltage, due to the inductance L.

The amount $|R_L|$ of the inductive resistance is defined as the ratio

$$|R_L| = \frac{U_0}{I_0} = \omega \cdot L \tag{5.17}$$

If the phase shift is taken into account the phase shifting resistor can be expressed by the complex number Z. It can be illustrated in a complex plane (x.iy) in Fig. 5.27 pointing into the imaginary axis iy. Its amount is $|Z| = |R_L|$ and its angle against the x-axis equals the phase shift φ between voltage and current [2].

It is tan $\varphi = \mathrm{Im}(Z)/\mathrm{Re}(Z)$ (see Vol.1 Sect. 13.3.2). The real part of Z is zero. This means that the inductance does not consume in the average any power.

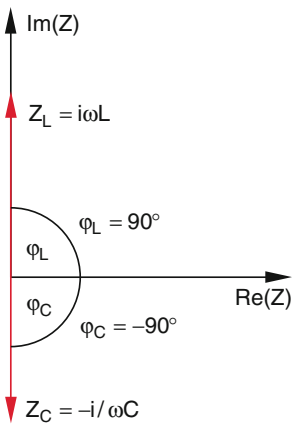


Fig. 5.27 Complex representation of inductive and capacitive resistances

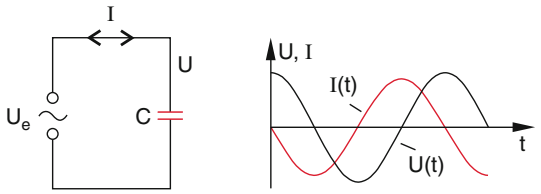


Fig. 5.28 AC circuit with capacitance C

5.4.2 Circuit with Capacitance

From the equation

$$U = Q/C$$

follows by differentiation with respect to time

$$\frac{dU}{dt} = \frac{1}{C}\frac{dQ}{dt} = \frac{1}{C} \cdot I. \tag{5.18a}$$

With $U_e = U_0 \cdot \cos \omega t$ we get

$$I = -\omega C \cdot U_0 \cdot \sin \omega t = \omega C \cdot U_0 \cdot \cos(\omega t + 90°). \tag{5.18b}$$

In a circuit with a capacitance C the current $I(t)$ is ahead of the voltage $U(t)$ by 90° (Fig. 5.28).

The complex resistance of the capacitance C is with $I_0 = \omega\ C\ U_0$

$$Z = \frac{U}{I} = e^{-i\pi/2}\frac{U_0}{I_0} = -i\frac{1}{\omega C} = \frac{1}{i\omega C}. \tag{5.19}$$

5.4.3 General Case

We now consider an ac-circuit that includes an ohmic resistor R, an inductance L and a capacitor C which are connected in series. With the external voltage $U_e(t)$ the sum of external voltage and induced voltage U_{ind} in Fig. 5.29 must be equal to

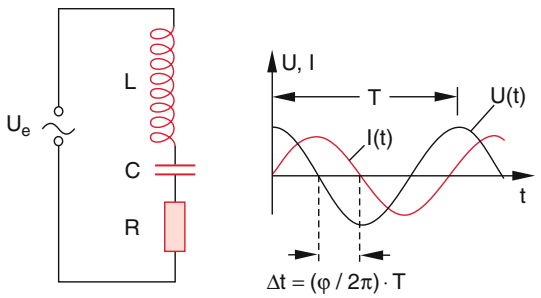


Fig. 5.29 General case of a series ac-circuit with inductance L, capacitance C and Ohmic resistor R

the voltage $U_1 + U_2 = I \cdot R + Q/C$ across the resistor R and the capacitor C. We therefore have the condition

$$U_e = L \cdot \frac{dI}{dt} + \frac{Q}{C} + I \cdot R. \tag{5.20}$$

Differentiation with respect to time gives

$$\frac{dU_e}{dt} = L \cdot \frac{d^2I}{dt^2} + \frac{1}{C} I + R \cdot \frac{dI}{dt}. \tag{5.21}$$

We try the complex solution

$$U_e = U_0 \cdot e^{i\omega t}, \quad I = I_0 \cdot e^{i(\omega t - \varphi)}. \tag{5.22}$$

Every reasonable solution must be, of course, real. For the solution we use the following properties of linear differential equations:

If the functions $f(t)$ and $g(t)$ are solutions of (5.21) then any linear combination $af(t) + bg(t)$ is also a solution, in particular the complex function $U(t) = f(t) + i\,g(t)$ This implies: When we have found a complex solution $U(t)$, the real part as well as the imaginary part are both solutions of (5.21). The special solution is determined by the initial conditions. (see Vol. 1, Chap. 11).

The complex ansatz allows a more simple notation and in particular a more elegant way to find the solution. Inserting (5.22) into (5.21) gives for the relation between current and voltage

$$i\,\omega\,U = \left(-L\omega^2 + i\omega R + 1/C\right) I. \tag{5.23}$$

When we define in analogy to the Ohmic resistor R the complex resistor Z by $Z = U/I$ we obtain from (5.23):

$$Z = \frac{U}{I} = R + i\left(\omega L - \frac{1}{\omega C}\right). \tag{5.24}$$

The complex resistor can be visualized by a vector in the complex plane (Fig. 5.27). Its amount

$$|Z| = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \tag{5.25}$$

Is called **impedance.**

The phase shift φ caused by the complex resistor Z is described by the ratio

$$\tan\varphi = \frac{\mathrm{Im}\{Z\}}{\mathrm{Re}\{Z\}} = \frac{\omega L - \frac{1}{\omega C}}{R} \tag{5.26}$$

of imaginary and real part. In the polar representation (see Vol. 1. Sect. 13.3.2)

$$Z = |Z| \cdot e^{i\varphi}$$

$|Z|$ gives the length of the vector and φ the angle against the x-axis.

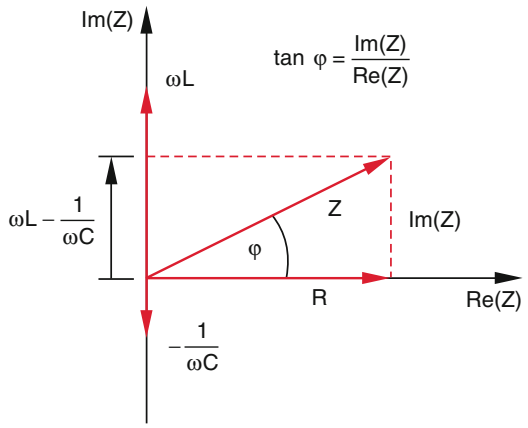


Fig. 5.30 Representation of the total resistance Z in the complex plane

The representation of complex resistors as vectors in the complex plane is called in electrical engineering a vector diagram. We will illustrate its usefulness by several examples in the next section.

From Fig. 5.30 and from Eq. (5.24) we see that for

$$\omega L = \frac{1}{\omega C}$$

The imaginary part of Z is zero. This implies that the phase shift between current and voltage becomes zero. It is therefore possible to make the idle power in a circuit with inductances and capacitors zero by a proper choice of the two phase shifting elements.

The current $I(t)$ through the ac circuit in Fig. 5.29 with the external voltage

$$U(t) = U_0 \cos\omega t$$

can be written as

$$I(t) = I_0 \cos(\omega t - \varphi)$$

With

$$I_0 = \frac{U_0}{\sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}}; \tag{5.27a}$$

and

$$\tan\varphi = \frac{\omega L - \frac{1}{\omega C}}{R}. \tag{5.27b}$$

The tangent of the phase shift φ between current and voltage is equal to the ratio of imaginary and real part of the complex resistance of a circuit (Fig. 5.30).

5.5 Linear Networks; High- and Low Frequency Passes; Frequency Filters

Linear networks are characterized by the linear relation between current and voltage, i.e.

$$U = Z \cdot I \tag{5.28}$$

This can be considered as the complex form of Ohm's Law $U = I \cdot R$ (2.6a)

If several currents with different frequencies are present in a linear network, it is possible to determine the currents $I(\omega_i)$ from the voltages $U(\omega_i)$ for any of the frequencies ω_i. The total current is then the sum of all currents $I(\omega_i)$.

This superposition principle which follows from the linearity of the network can be expressed in a complex notation as

$$\begin{aligned} U(t) &= \sum_k U_k(\omega_k) \\ &= \sum_k U_{0k} e^{i(\omega_k t - \varphi_k)}, \end{aligned} \tag{5.29a}$$

$$\begin{aligned} I(t) &= \sum_k I_{0k} e^{i(\omega_k t - \psi_k)}, \\ \Rightarrow Z_k(\omega_k) &= \frac{U_{0k}}{I_{0k}} \cdot e^{i(\psi_k - \varphi_k)}. \end{aligned} \tag{5.29b}$$

The superposition principle is of great importance for high frequency technology since it allows the determination of complex voltage or current pulses and their changes when these pulses pass through linear networks. The input pulses are decomposed into their frequency components $U_e(\omega)$ and $I_e(\omega)$ (Fourier analysis). For each frequency component the change of amplitude and phase is calculated when it passes through the network and finally all modified output frequency components $U_a(\omega)$ and $I_a(\omega)$ are added again to obtain the total final output pulse (Fourier Synthesis). This shall be illustrated by some examples [3]:

5.5.1 High-Frequency Pass

An electrical high pass is a circuit, that lets pass all high frequencies ω barely attenuated, but blocks all low frequencies. Figure 5.31 shows one of several realizations. The input voltage $U_e(t) = U_0 \cos \omega t$ is reduced by the frequency-dependent voltage divider to the output voltage

$$U_a = \frac{R}{R + \frac{1}{i\omega C}} \cdot U_e. \tag{5.30a}$$

Multiplying numerator and denumerator with the conjugate complex of the denumerator yields

$$U_a = \frac{R^2\omega^2C^2 + iR\omega C}{1 + \omega^2R^2C^2} \cdot U_e. \tag{5.30b}$$

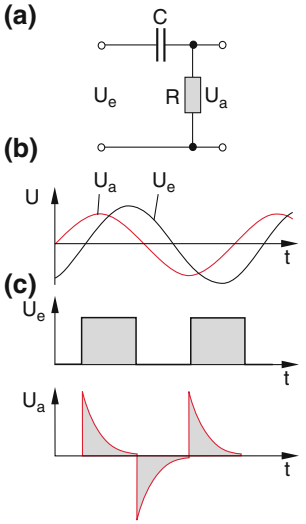


Fig. 5.31 High frequency pass: **a**) circuit, **b**) input and output voltage for a cosine voltage, **c**) for a square-wave input voltage

$$|U_a| = \frac{\omega \cdot R \cdot C}{\sqrt{1 + \omega^2R^2C^2}} \cdot |U_e|. \tag{5.31}$$

For the phase shift between output and input voltage we obtain

$$\tan \varphi = \frac{1}{\omega RC}. \tag{5.32}$$

From Eq. (5.31) we see, that for a high frequency pass the ratio $|U_a|/|U_e| = 0$ for $\omega = 0$ and increases for increasing ω. For $\omega = 1/RC$ it has the value $1/\sqrt{2}$ and for $\omega \to \infty$ it becomes 1 (Fig. 5.29). The phase shift φ drops from 90° at $\omega = 0$ to 0° for $\omega \to \infty$.

It is interesting to study the transmission of a rectangular pulse through a high frequency pass. The Fourier analysis of a regular sequence of rectangular pulses shows that the steep edges of the pulse correspond to the high frequencies whereas the flat roof is represented by the low frequencies. Since the high pass attenuates the high frequencies much less than the low frequencies, the rising and falling edges of the pulse barely decrease during the transmission through the high pass.

Another way to understand this, is the following:

The sudden voltage jump at the left capacitor plate in Fig. 5.31c is transferred by influence onto the right plate, which is discharged through the resistor R with the time constant $\tau = R \cdot C$.

The voltage across the capacitor is $U = Q/C$, the output voltage $U_a = I \cdot R$. With $I = dQ/dt$ we get the output voltage

$$U_a = \frac{dQ}{dt} \cdot R = R \cdot C \cdot \frac{dU_e}{dt}. \tag{5.33}$$

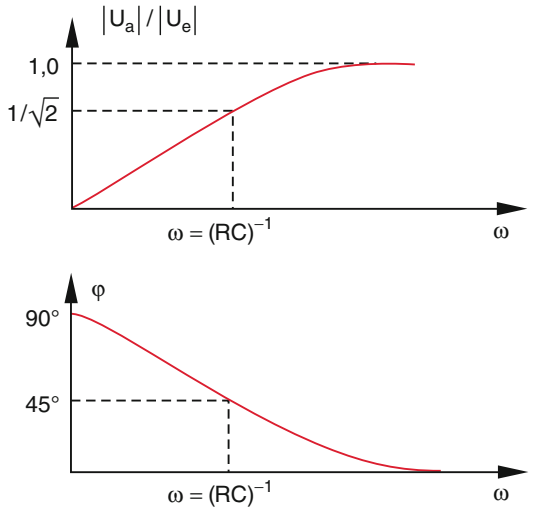


Fig. 5.32 Ratio $|U_a|/|U_e|$ of the amplitudes of output- to input voltages and phase shift between output- and input voltage for a high frequency pass

The output voltage U_a is proportional to the time derivative of the input voltage U_e. Therefore the high pass is also called differentiating element, which is used in analogue computers to perform the mathematical operation of differentiation (Fig. 5.31c).

5.5.2 Low Frequency Pass

For the example of a low pass shown in Fig. 5.33 resistor R and capacitance C are just interchanged compared to the high pass in Fig. 5.31a. From Fig. 5.33 we can directly obtain the equation of the voltage divider consisting of R and C in series.

$$\begin{aligned} U_a &= \frac{1/(i\omega C)}{R + 1/(i\omega C)} \cdot U_e \\ &= \frac{1}{1 + i\omega RC} \cdot U_e, \end{aligned} \tag{5.34}$$

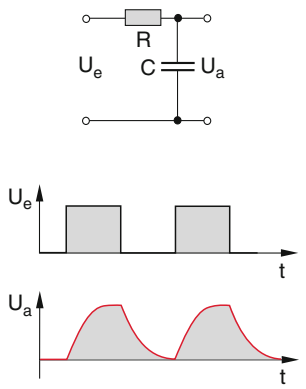


Fig. 5.33 Low frequency pass (integration circuit)

This gives for the amounts of input and output voltage

$$|U_a| = \frac{1}{\sqrt{1 + \omega^2 R^2 C^2}} \cdot |U_e|, \tag{5.35a}$$

And for the phase shift

$$\tan \varphi = -\omega RC. \tag{5.35b}$$

For a low frequency pass the ratio $|U_a|/|U_e|$ decreases from 1 at $\omega = 0$ to zero at $\omega = \infty$.

The output voltage

$$\begin{aligned} U_a &= \frac{Q}{C} = \frac{1}{C} \int I \, dt \\ &= \frac{1}{RC} \int (U_e - U_a) \, dt \end{aligned} \tag{5.36}$$

Is proportional to the integral over the difference $U_e - U_a$. Therefore the low frequency pass is called an integration element and is used in analogue computers to perform the mathematical operation of integration (Fig. 5.33, lower part).

5.5.3 Frequency Filters

The circuit in Fig. 5.29 can be used as RCL bandpass filter. This can be seen, when we determine the output voltage of the circuit in Fig. 5.34a. We get:

$$U_a = \frac{R}{R + i\left(\omega L - \frac{1}{\omega C}\right)} \cdot U_e \tag{5.37}$$

$$|U_a| = \frac{R}{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} \cdot |U_e|. \tag{5.38}$$

For the resonance frequency

$$\omega = \omega_R = \frac{1}{\sqrt{L \cdot C}} \tag{5.39}$$

is $|U_a| = |U_e|$. The ac-voltage $U_e(\omega_R)$ at the resonance frequency ω_R is transmitted through the frequency filter without any attenuation.

For $\omega L - 1/(\omega C) = \pm R$ the output voltage U_a drops to $U_e/\sqrt{2}$. This gives the condition

$$\omega_{1,2} = \pm \frac{R}{2L} + \sqrt{\frac{R^2}{4L^2} + \omega_R^2}, \tag{5.40}$$

for the two frequencies ω_1 and ω_2 where U_a has dropped to $U_e/\sqrt{2}$. The width of the transmission curve (Fig. 5.34b)

$$T(\omega) = |U_a|/|U_e|$$

is

$$\Delta\omega = \omega_1 - \omega_2 = \frac{R}{L} \tag{5.41}$$

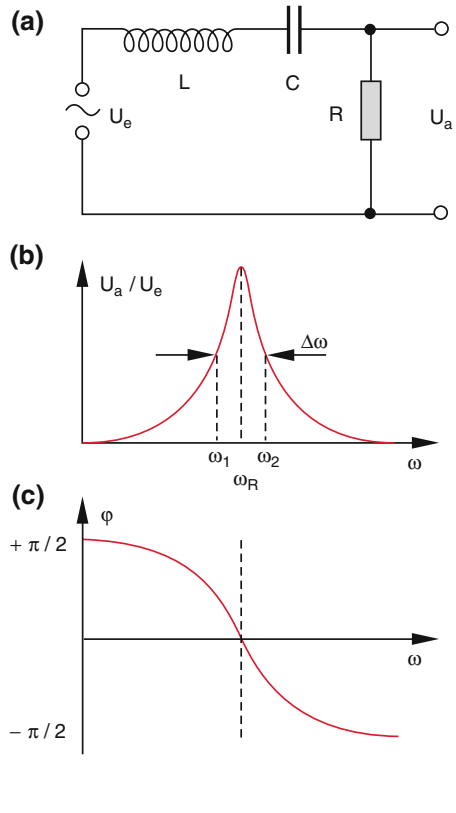


Fig. 5.34 Frequency dependent transmission filter. **a**) Circuit, **b**) transmission curve T(ω), **c**) phase shift $\varphi(\omega)$ between U_a and U_e

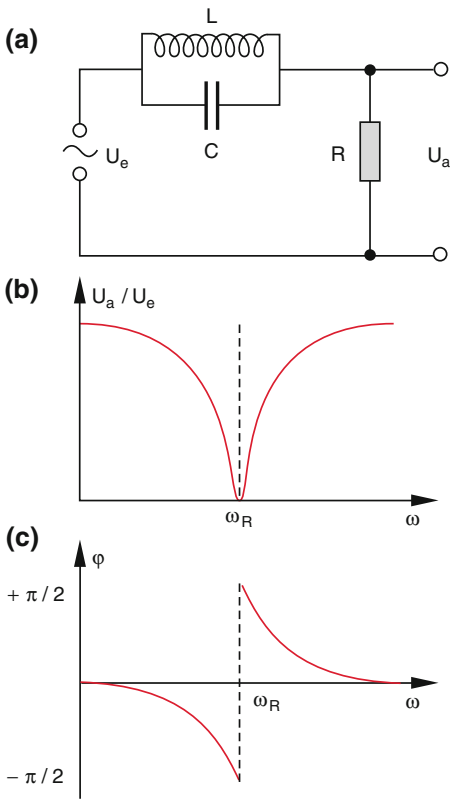


Fig. 5.35 Frequency dependent blocking filter. **a**) = circuit, **b**) transmission curve T(ω), **c**) phase shift $\varphi(\omega)$ between U_a and U_e

The output voltage is delayed against the input voltage. The phase shift φ between output and input voltage in the circuit of Fig. 5.34a is shown in Fig. 5.34c. It is given by

$$\tan\varphi = \frac{1/\omega C - \omega L}{R}. \tag{5.42}$$

For $\omega = 0$ is $\varphi(0) = +90°$, for $\omega = \omega_R$ is $\varphi(\omega_R) = 0$ and it becomes $-90°$ for $\omega = \infty$, $\varphi(\infty) = -90°$.

While the circuit in Fig. 5.34a has its *maximum* transmission at $\omega = \omega_R$ the circuit in Fig. 5.35a (blocking filter) has its *minimum* transmission at the resonance frequency ω_R. It is $U_a(\omega_R) = 0$.

The transmission curves $T(\omega)$ are shown in Fig. 5.34b for the transmission bandpass filter and in Fig. 5.35b for the blocking filter.

Compare Fig. 5.34 with the completely similar Fig. 11.22 in Vol. 1 for the forced oscillation. Explain this similarity!

5.6 Tranformers

In order to transport the electric power $P_{el} = U \cdot I$ over large distances, it is advantageous to minimize the transmission losses by Joule's heat power $\Delta P_{el} = I^2 \cdot R$, due to the resistance R of the transmission lines. For a given transmitted power P_{el} the losses decrease with decreasing current. One should therefore choose the voltage U as high as possible in order to minimize the current $I = P_{el}/U$.

The relative transmission losses

$$\frac{\Delta P_{el}}{P_{el}} = \frac{I^2 \cdot R}{U \cdot I} = \frac{I \cdot R}{U} = \frac{R}{U^2} P_{el} \tag{5.43a}$$

Decrease as $1/U^2$ with increasing voltage. The resistance R of the transmission line causes a voltage drop $\Delta U = I \cdot R$. From (5.43a, 5.43b) we then obtain

$$\frac{\Delta P_{el}}{P_{el}} = \frac{\Delta U}{U}. \tag{5.43b}$$

Example

A copper cable with 2.5 km length and a cross section of 0.2 cm^2 has at the temperarture T = 20 °C a specific resistance $\rho_{el} = 1.7 \cdot 10^{-8} \Omega \cdot m$ and therefore a total resistance of $R = 2.1\,\Omega$. If a power of $P_{el} = 20\,kW$ should be transmitted at a voltage of $U = 230\,V$ a

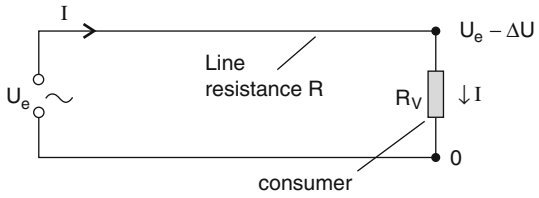


Fig. 5.36 Schematic illustration of the power loss in power lines

current of $I = 87$ A is needed. The voltage drop along the copper cable is, however, already $\Delta U = I \cdot R = 185$ V and therefore the voltage at the consumer is only 45 V. The relative power loss is $\Delta P_{el}/P_{el} = 0.80$. This means that only 20% of the original power provided by the generator arrives at the consumer!

If, however, the voltage is transformed up to 20 kV, only a current of 1 A is needed for the same transmitted power. The voltage drop is now only $\Delta U = 2.1$ V and the relative power loss in the transmission line is $\Delta P_{el}/P_{el} = 10^{-4}$.

This example demonstrates that for sufficiently high voltages the transmission losses are (opposite to the common opinion), negligible (Fig. 5.36). For instance the transmission of $P_{el} = 10$ MW over a distance of 300 km at a voltage of 380 kV demands a current $I = 26$ A and causes a voltage drop in the transmission line with a resistance of 0.3 Ω/km of $\Delta U = 0.3 \cdot 300 \cdot 26$ V $= 2.4$ kV and a relative power loss $\Delta P_{el}/P_{el} = 2.4/380 = 0.62\%$ (Fig. 5.37)

The transformation of voltages is realized with transformers (Fig. 5.39). Their principle is based on Faraday's induction law.

Two coils L_1 and L_2 with the number N_1 and N_2 of windings are coupled by an iron yoke in such a way, that the magnetic flux generated in coil 1 (primary coil) passes completely through coil 2 (secondary coil) (Fig. 5.38b). Due to the large magnetic permeability μ of iron all magnetic field lines generated in coil 1 pass through the iron core of coil 2. In order to avoid eddy currents in the iron yoke, which would result in heat losses, the yoke consists of many thin sheets of iron, which are isolated against each other by a thin isolating layer. They are pressed tightly together by isolated screws to avoid vibrations of the sheets induced by the alternating magnetic field at the frequency $2\nu = \omega/\pi$, which exert forces onto the sheets which results in an annoying noise (transformer drone).

5.6.1 Tranformer Without Load

We will at first consider the tranformer without load where no current flows through the secondary coil ($I_2 = 0$).

Fig. 5.37 Installation of a high voltage power line. In this picture each phase of the three phase current is transmitted by 4 cables arranged in quadratic form, connected by the diagonal bracket seen in the upper part of the picture. Each cable consists of two wires. The technician just fixes the connection between these wires (with kind permission by information center of the Elektrizitätswirtschaft Frankfurt)

When an external ac-voltage

$$U_1 = U_0 \cos \omega t$$

is applied to the primary coil with inductance L_1 and N_1 windings the current I_1 flows through L_1 and induces a voltage

$$U_{ind} = -L_1 \frac{dI_1}{dt} = -N_1 \frac{d\Phi_m}{dt} = -U_1, \qquad (5.44a)$$

which is opposite to the external voltage U_1, since according to Kirchoff's rule the total voltage in a closed circuit must be zero:

$$U_1 + U_{ind} = 0. \qquad (5.44b)$$

Here we have neglected the ohmic resistance of the inductance, because it is very small compared to ωL. If the total magnetic flux Φ_m, generated in the primary coil L_1 passes through the secondary coil L_2 with N_2 windings the voltage, induced in L_2 is

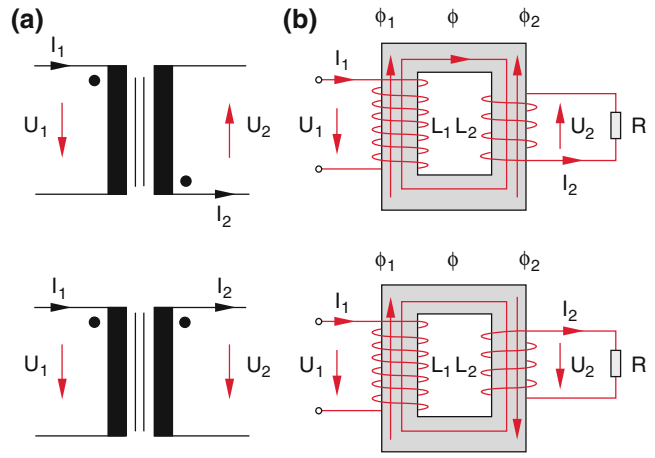


Fig. 5.38 Transformator **a**) schematic circuit, **b**) technical design. In the upper part primary and secondary windings have the same direction of windings in the lower part opposite directions. In the upper part the output voltage is shifted by 180° against the input voltage in the lower part they are in phase. This is indicated by the black points in (**a**)

$$U_2 = -N_2 \frac{\mathrm{d}\Phi_\mathrm{m}}{\mathrm{d}t} \tag{5.45}$$

with $\mathrm{d}\Phi_\mathrm{m}/\mathrm{d}t = U_1/N_1$ it follows from (5.45) and (5.44a, 5.44b)

$$\frac{U_2}{U_1} = -\frac{N_2}{N_1}. \tag{5.46}$$

The minus sign indicates, that U_1 and U_2 have the opposite sign, (i.e. U_2 is phase shifted by 180° against U_1) if the windings of the two coils are both wound in the same direction (Fig. 5.38 above). The two voltages are in phase, if the windings are in the opposite direction (Fig. 5.38 below). For $N_2 > N_1$ the secondary voltage is larger than the input voltage (up-transformation).

The average power of the transformer without load (lossless coils, no load connected to the secondary coil) is

$$\bar{P}_\mathrm{el} = \frac{1}{2} U_{01} I_{01} \cos\varphi \equiv 0, \tag{5.47}$$

Because the phase shift φ between current and voltage is, according to (5.16), $\varphi = 90°$. The current in the primary coil is a pure idle current and does not consume energy [4].

5.6.2 Transformer with Load

When the secondary coil is connected to a load with resistance R, the current through the coil is $I_2 = U_2/R$. This current produces a magnetic flux $\Phi_2 \propto I_2$ which is phase-shifted by 90° against the flux Φ_1. The current I_2 is in phase with $U_2 = RI_2$ but the phase of the voltage $U_2 = -N_2 \mathrm{d}\Phi_2/\mathrm{d}t$ is shifted by 90° against Φ_1.

This magnetic flux Φ_2 generated by the current I_2 is superimposed to the flux Φ_1 and gives the total flux

$$\Phi = \Phi_1 + \Phi_2,$$

which has the phase shift $\Delta\varphi$ $0 < \Delta\varphi < 90°$ against the input voltage U_1 with $\tan\Delta\varphi = \Phi_2/\Phi_1$

This superposition of the magnetic fluxes has the consequence that in addition to the current I_1 a second phase-shifted contribution caused by the magnetic flux Φ_2 superimposes the current in the primary coil. Now the power consumed by the primary coil is no longer a pure eddy power, but includes aln effective power. The total mean power received by the transformer is now

$$\bar{P}_\mathrm{el} = \frac{1}{2} U_0 \sqrt{I_{01}^2 + I_{02}^2} \cdot \cos(\varphi - \Delta\varphi) \tag{5.48}$$

It is no longer zero, because $\varphi - \Delta\varphi \neq 90°$.

The quantitative description of the ideal transformer with a secondary side connected to a complex resistance Z starts from the equations

$$U_1 = \mathrm{i}\omega L_1 I_1 + \mathrm{i}\omega L_{12} I_2, \tag{5.49a}$$

$$U_2 = Z \cdot I_2 = -\mathrm{i}\omega L_{12} I_1 - \mathrm{i}\omega L_2 I_2, \tag{5.49b}$$

where we have neglected any heat losses in the coils or the iron core and also all transmission losses of the magnetic flux. The quantities L_1 and L_2 are the inductances of the primary and secondary coil and L_{12} is the mutual inductance. The voltage U_1 generates the current I_2 and is ahead of the current by 90°, the induced voltage U_2, however, is in phase with the current if the load is a resistor R, but is phase-shifted

for a complex load Z which induces a phase shift between voltage and current (see below). This is indicated in (5.49a, 5.49b, 5.49c) by the minus sign.

For the lossless transformer the input power equals the output power, i.e.

$$I_1 \cdot U_1 = I_2 \cdot U_2. \tag{5.49c}$$

Solving (5.49b) for I_2 and inserting this into (5.49a) gives the relations between the currents I_1, I_2 and the input voltage U_1

$$I_1 = \frac{\mathrm{i}\omega L_2 + Z}{\mathrm{i}\omega L_1 Z + \omega^2(L_{12}^2 - L_1 L_2)} \cdot U_1, \tag{5.50a}$$

$$I_2 = -\frac{\mathrm{i}\omega L_{12}}{\mathrm{i}\omega L_1 Z + \omega^2(L_{12}^2 - L_1 L_2)} \cdot U_1. \tag{5.50b}$$

This yields the ratio of output to input current

$$\frac{I_2}{I_1} = -\frac{\mathrm{i}\omega L_{12}}{\mathrm{i}\omega L_2 + Z} \tag{5.51}$$

With $I_2 = U_2/Z$ we obtain the ratio of output to input voltage

$$\frac{U_2}{U_1} = -\frac{\mathrm{i}\omega L_{12} Z}{\mathrm{i}\omega L_1 Z + \omega^2(L_{12}^2 - L_1 L_2)}. \tag{5.52a}$$

The strength of the magnetic coupling k between primary and secondary coil is defined by

$$k = \frac{L_{12}}{\sqrt{L_1 \cdot l_2}} \quad \text{with} \quad 0 < k < 1$$

For a complete coupling (no coupling losses) is $k = 1$; i.e. $L_{12} = \sqrt{L_1 L_2}$) For $k = 1$ (5.52a) reduces to

$$\frac{U_2}{U_1} = \frac{L_{12}}{L_1 - \mathrm{i}\omega(k^2 - 1)L_1 L_2/Z}. \tag{5.52b}$$

For the amounts we obtain

$$\frac{|U_2|}{|U_1|} = \frac{L_{12}/L_1}{1 + (\omega^2 L_2^2/|Z|^2)(k^2 - 1)^2}. \tag{5.52c}$$

We will now discuss transformers with special loads $Z = R$ (pure ohmic load), $Z = L$ (pure inductive load) and $Z = C$ (pure capacitive load).

5.6.2.1 Z = R

For a complete coupling ($k = 1$ and $L_{12} = \sqrt{L_1 \cdot L_2}$) we obtain instead of (5.52c)

$$\frac{U_2}{U_1} = \frac{L_{12}}{L_1} = -\sqrt{\frac{L_2}{L_1}} = -\frac{N_2}{N_1}, \tag{5.53}$$

Because according to (4.10) is $L \sim N^2$.

In case of complete coupling the ratio U_2/U_1 is independent of the load resistor R.

This is, however, only valid if the resistance of the transformer coils and therefore the voltage drop across them are negligible. This was anticipated in (5.49a, 5.49b, 5.49c).

For $k < 1$ the ratio $|U_2|/|U_1|$ decreases with decreasing R (see 5.52d). This implies that with increasing current load the secondary voltage U_2 drops.

Example

With $k = 0.9$ the ratio U_2/U_1 drops for $R = 0.1 \cdot |\omega L_2|$ to $1/\sqrt{2} = 0.71$ of its value for the transformer without load.

The phase shift φ between U_2 and U_1 can be obtained from (5.52c) as

$$\tan\varphi = -\frac{\omega L_2(1 - k^2)}{R}. \tag{5.54}$$

For $k \to 1$ the phase shift $\varphi \to 180°$ independent of R. For incomplete coupling ($k < 1$) the phase shift φ becomes $\varphi(k < 1) < 180°$.

5.6.2.2 Z = iωL (Pure Inductive Load)

From (5.52b) and (5.52c) we obtain the ratio

$$\frac{U_2}{U_1} = -\frac{L_{12}/L_1}{1 + (L_2/L)(1 - k^2)}. \tag{5.55}$$

This ratio is real in spite of the imaginary load $\mathrm{i}\omega L$. The phase shift is always $\varphi = 180°$. The voltage ratio depends on the expression $(\mathrm{L_2/L}) \cdot \left(1 - \mathrm{k}^2\right)$.

Example

With $k = 0.9$ and $L_2/L = 10$ we obtain from (5.55). $(U_2/U_1)_L = \frac{1}{2}(U_2/U_1)_{L=\infty}$. Where $L = \infty$ corresponds to the case with no load. The output voltage drops to ½ of the case with no load. This can be understood because the parallel circuit of $L = 0.1\,L_2$ to the additional load of the secondary coil has the same effect as the coupling losses of 10% for $k = 0.9$. For $k = 1$ the load L does not affect the output voltage U_2.

5.6.2.3 $Z = 1/(i\omega C)$ (Pure Capacitive Load)

The ratio

$$\frac{U_2}{U_1} = \frac{L_{12}}{L_1 - \omega^2 C L_1 L_2 (1 - k^2)} \tag{5.56a}$$

becomes *larger*!! than for the transformer with no load ($Z = \infty$) described by (5.52c). For the resonance frequency

$$\omega_R = \sqrt{\frac{1}{C L_2 (1 - k^2)}} \tag{5.56b}$$

U_2 becomes infinite if all losses in the transformer can be neglected.

This is called the *resonant voltage step up* of the transformer output voltage.

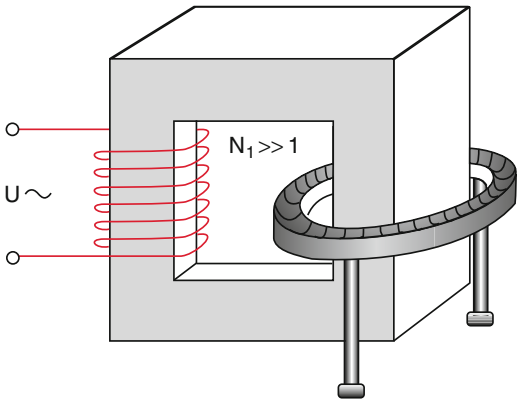


Fig. 5.40 Transformer with a single secondary winding used for melting of metals

5.6.3 Applications

Transformers play an important role for many technical applications. They transform ac-voltages to higher or lower values and they are indispensable for the generation of very high ac- currents. For the transformation from the high voltage lines to the medium voltage networks that distributes the electrical power to the different neighborhoods of a city, transformer stations have been constructed with many transformers (Fig. 5.39a). One of these transformers is shown in Fig. 5.39b.

An example for the transformation to high currents (i.e. lower voltages) is shown in Fig. 5.40. The secondary coil consists of only one winding, which is formed as a gutter. With the current I_2 the power $I_2^2 \cdot R$, dissipated in the gutter, can become so large that solid metal in the gutter with a lower melting point than the gutter itself becomes liquid. Such high current transformers are used for melting aluminum in the aluminum producing industry.

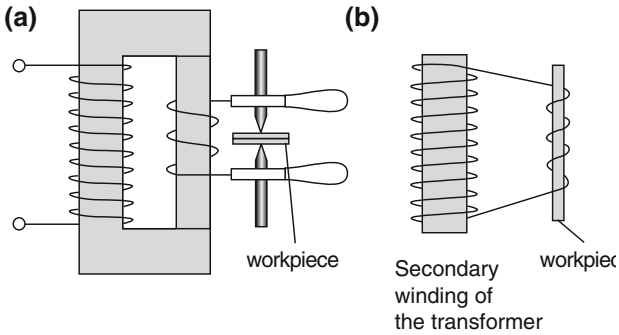


Fig. 5.41 Transformer with high secondary current **a**) used for spot welding, **b**) for heating up a metal rod by eddy currents

Example

$N_1 = 100$, $N_1 = 1$; $U_1^{\text{eff}} = 230\,\text{V}$, $R = 5 \cdot 10^{-3}\,\Omega$,
$\rightarrow U_2^{\text{eff}} = 2.3\,\text{V} \rightarrow I_2^{\text{eff}} = 460\,\text{A}$; $\bar{P}_{\text{el}} = I_2^2 R = 1.06\,\text{kW}$.

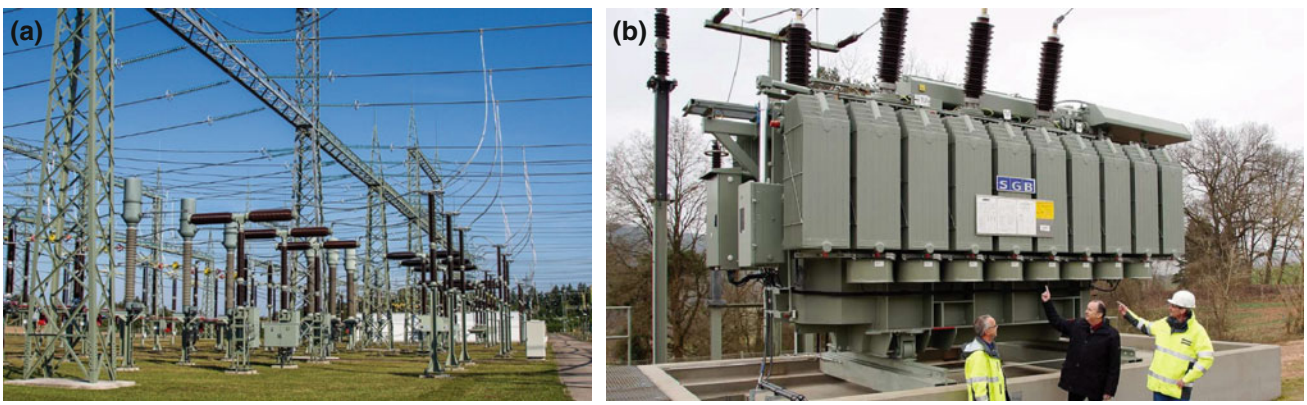


Fig. 5.39 **a**) High voltage transformator field, **b**) high voltage transformer station

The large current can be also used for spot welding (Fig. 5.41a). The two sharpened pins form part of the secondary coil. The two work pieces are brought between the two pins which can be pressed against each other with isolated handles. The secondary current flows through a small spot of the two work pieces which melt and are welded together.

If the secondary coil of the transformer with only a few windings is connected to another coil around a metallic rod (Fig. 5.41b) the high ac current heats the rod due to inductive heating to such high temperatures that it glows bright red. Many electro stoves use this induction principle for heating the metallic bottom of cooking pots.

Most electronic devices demand a low voltage for their power supply. Transformers can provide any wanted voltage by choosing the appropriate ratio N_1/N_2.

Small high voltage transformers with a high ratio N_2/N_1 and secondary voltages of 10–20 kV provide the necessary high voltage in older TV devices for the deflection of the electron beam. They are still used in X-ray tubes.

5.7 Impedance Matching in ac-Circuits

Often the problem arises to transfer the maximum power from the source to an electric circuit with a complex resistance Z. This is only possible if the complex resistances of source and load are matched. In order to find the optimum matching conditions we consider in Fig. 5.42 an ac- source with the voltage $U = U_0 \cos \omega t$ which is connected with the load resistor Z_2 by a "matching resistor" Z_1. From Fig. 5.42 we obtain

$$Z_2 = R_2 + \mathrm{i}\left(\omega L_2 - \frac{1}{\omega C_2}\right) \tag{5.57a}$$

The effective current through the whole circuit is

$$I_{\mathrm{eff}} = U_{\mathrm{eff}}/Z \quad \text{with} \quad Z = Z_1 + Z_2.$$

The real power consumed in the load is

$$\bar{P}_{\mathrm{el}} = I_{\mathrm{eff}}^2 \cdot R_2 = \frac{U_{\mathrm{eff}}^2}{|Z|^2} \cdot R_2. \tag{5.57b}$$

Inserting the complex resistance

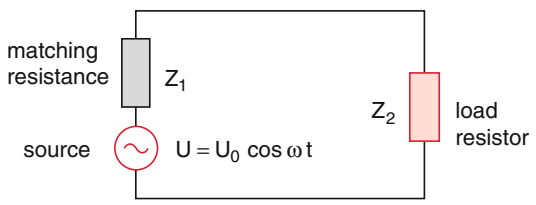


Fig. 5.42 Adapting the load resistance Z_2 to the complex source resistance Z_1 for optimum power transfer

$$Z = R_1 + R_2 + \mathrm{i}\left[\omega(L_1 + L_2) - \frac{1}{\omega}\left(\frac{1}{C_1} + \frac{1}{C_2}\right)\right] \tag{5.57c}$$

We obtain the real power consumption

$$\bar{P}_{\mathrm{el}} = \frac{U_{\mathrm{eff}}^2 \cdot R_2}{(R_1 + R_2)^2\left[\omega(L_1 + L_2) - \frac{1}{\omega}\left(\frac{1}{C_1} + \frac{1}{C_2}\right)\right]^2}. \tag{5.58}$$

From (5.58) we conclude immediately that $\bar{P}_{\mathrm{el}}$ becomes maximum, if the second bracket in the denominator becomes zero, i.e. if

$$\omega L_2 - \frac{1}{\omega C_2} = -\left(\omega L_1 - \frac{1}{\omega C_1}\right). \tag{5.59}$$

With this condition $\bar{P}_{\mathrm{el}}$ depends only on R_1 and R_2. For a given value of R_1 the mean power consumption $\bar{P}_{\mathrm{el}}$ becomes maximum if $\mathrm{d}P_{el}/\mathrm{d}R_2 = 0$. This gives the condition $R_1 = R_2$.

Optimum power matching is achieved if the real resistances of source and load are equal. In this case no blind power is produced and the transferred real power becomes maximum.

5.8 Rectification

For many scientific and technical devices dc-voltages and currents are demanded. Therefore circuits have to be developed which transduce the ac-voltage supplied by the wall outlet or by the secondary coil of a transformer into a constant dc-voltage. This can be achieved with rectifiers using electron tubes or semiconductor devices.

The circuit symbols for different devices are shown in Fig. 5.43. The *technical current direction* (indicated by the arrow direction in the symbol) is defined (because of historical reasons) as the flow direction of positively charged particles, i. e from the anode (plus) to the cathode (minus), although we know today that the current in metals and semiconductors is carried by electrons with the opposite flow direction.

The rectifying diode opens (transmits the current) if a positive voltage is applied to the anode relative to the cathode. For negative voltages it blocks the current. A typical current-voltage characteristic of a rectifying diode is shown in Fig. 5.44. For small negative voltages only the small blocking current can pass through the diode, when the kinetic energy of the electrons emitted from the cathode can overcome the small negative voltage, i.e. if $E_{\mathrm{kin}} + e\,U > 0$ [5].

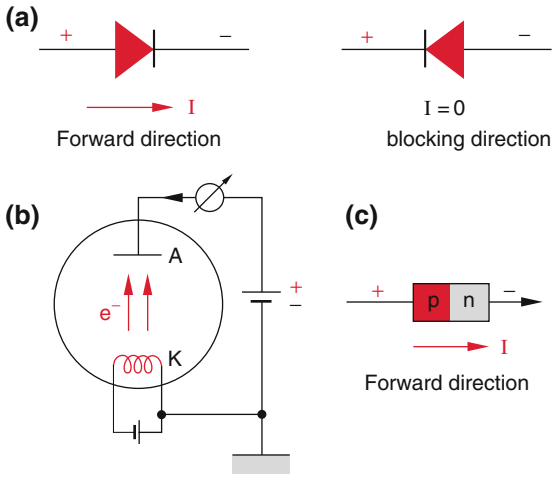


Fig. 5.43 **a**) Graphical symbol for a diode. The diode arrow points into the technical current direction which is opposite to the electron current. **b**) Thermionic diode in an evacuated glass bulb. **c**) Semiconductor diode

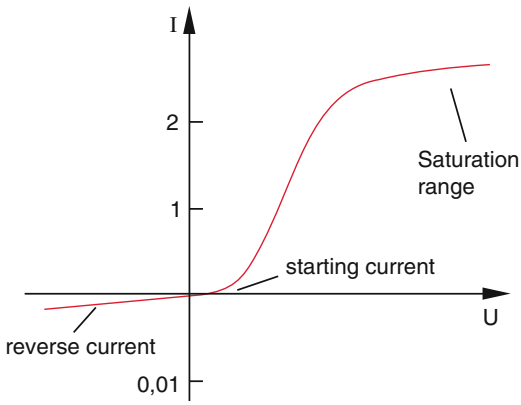


Fig. 5.44 Current-voltage characteristic of a diode with reverse current, starting current and saturation range

5.8.1 One-way Rectification

When only one diode is used (Fig. 5.45a) only the positive half of the ac-voltage can pass. This results in large ripples of the rectified voltage (Fig. 5.45b). The maximum dc-voltage is U_0. Even when using a smoothing capacitor (Fig. 5.46) the result is not satisfying for most applications which demand a smooth dc-voltage. The solution is the rectification with more than one diode.

5.8.2 Two-way Rectification

In the two-way rectification the midpoint of the secondary coil of the transformer represents the reference voltage, which is generally grounded, i.e. its voltage is zero. The two ends of the secondary coil are connected with two parallel

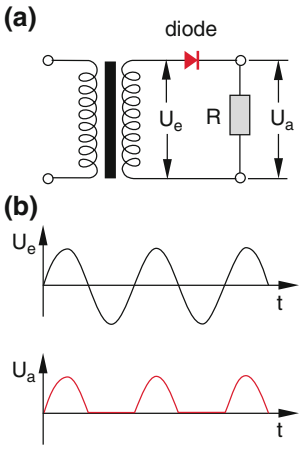


Fig. 5.45 One way rectification, **a**) circuit, **b**) comparison of the ac-voltage before and after rectification

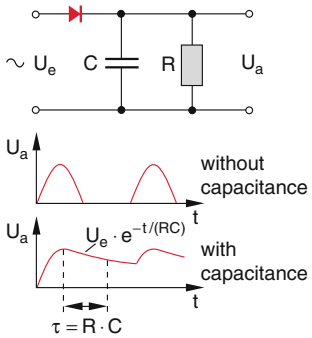


Fig. 5.46 Smoothening of the pulsating dc-current by a capacitor

diodes and the outputs of the two diodes are connected and form one pole of the dc-voltage (Fig. 5.47). The upper diode in Fig. 5.47 passes the current when the voltage between the upper end of the transformer coil and the lower end is positive, while the lower diode is open when this voltage is negative, i.e. when the voltage between midpoint and lower end of the coil is positive.

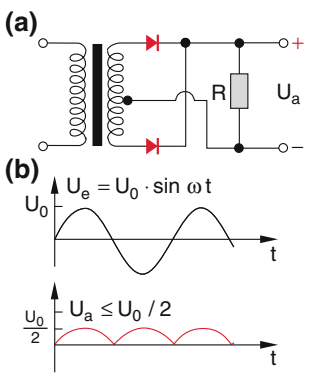


Fig. 5.47 Two-way rectification

This two-way rectification therefore overrides the gap of the one-way rectification (Fig. 5.47b).

The maximum dc-voltage is $U_0/2$ for an input ac-voltage $U = U_0 \cdot \cos \omega t$. A small disadvantage is that a midpoint tap of the secondary coil of the transformer is needed.

5.8.3 Bridge Rectifying Circuit

The mainly used rectifying circuit is the bridge circuit in Fig. 5.48, also called *Graetz- circuit*. It is supplied nowadays as small integrated semiconductor device for small and medium powers. From Fig. 5.48 one can see, that the same form of the output power as for the two-way rectification is obtained but with twice the dc-voltage. The smoothing of the pulsating dc voltage is optimized by the circuit shown in Fig. 5.49, consisting of a load capacitor C_1 and a frequency dependent voltage divider consisting of L and C_2.

The dc-output voltage for a voltage U_1 at the load capacitor C_1 is

$$U_a = \frac{U_1}{\sqrt{(1 - \omega^2 LC_2)^2 + \omega^2 L^2/R^2}}, \qquad (5.60)$$

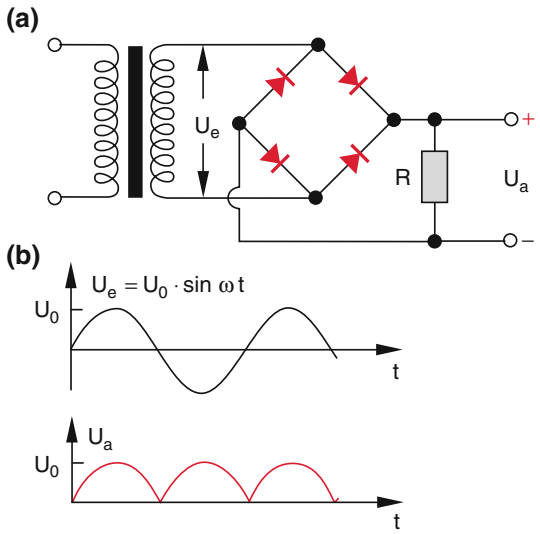


Fig. 5.48 Bridge rectification (Graetz-circuit)

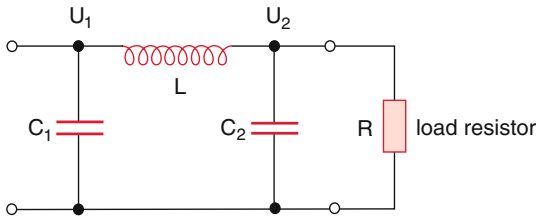


Fig. 5.49 Smoothing of the dc voltage with ripples by an L-C-circuit

This can be seen when considering the ac-resistances $i\omega L$ for the inductance L and $1/(i\omega C_2)$ for the capacitor C_2 for the parallel arrangement of C_2 and R. While the dc-voltage passes without attenuation the ac- contributions with $\omega > 0$ are diminished.

The filter circuit in Fig. 5.49 represents a special low pass filter.

Replacing the inductance by a resistor R one gets the low pass filter of Fig. 5.34. In the latter circuit, however, the dc-voltage is also attenuated (see Problem 5.9).

Example

$\Omega = 2\pi \cdot 50\,s\psi - 1\psi$, $R = 50\ \Omega$, $L = 1$ H, $C_2 = 10^{-3}$ F, $\rightarrow \omega L = 314\ \Omega$, $1/(\omega C) = 3\ \Omega$, $\rightarrow U_a(\omega) = 0.01\ U_e$, while for the dc-voltage is $U_a(\omega = 0) = U_e(\omega = 0)$.

In modern rectifying devices for small and medium powers (e.g. for power supplies of computers or for radios and television sets) the output voltage is electronically stabilized. This drops the ripples ΔU down to relative values $\Delta U/U < 10^{-3} - 10^{-4}$ [3].

For larger powers, the rotating three phase ac current (Fig. 5.50) is the best solution for obtaining smooth dc-currents with small ripples. Since the phase shift between the different phases is only 120°, the rectification by

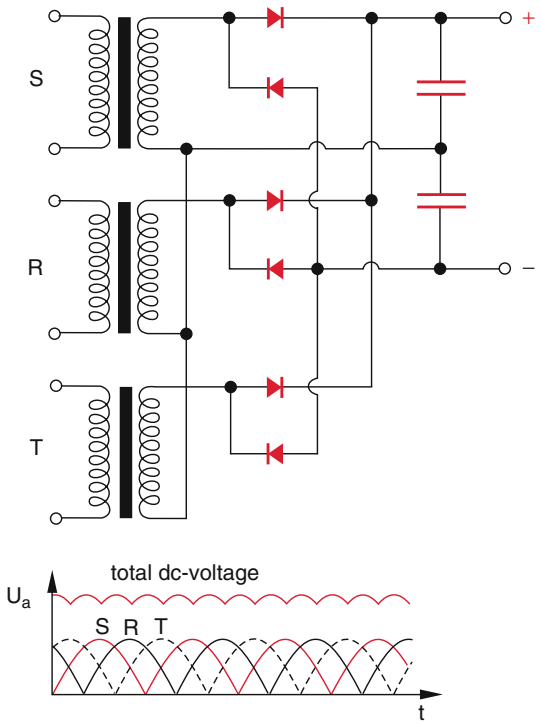


Fig. 5.50 Rectification of the three phases R, S, T of the three-phase ac voltage

two-way or bridge rectifiers, the dc-current has even without a load capacitor a ripple

$$\frac{U_{max} - U_{min}}{U_{max}} \approx 0.13 = 13\%$$

Compared to 100% for the two-way rectification of the one-phase ac-current. Using a load capacitor C the smoothing becomes much better. Between two peaks of the ac-voltage it drops only by 1/3 of that for the one-phase ac-current. Therefore the drop of the dc voltage $U(t) = U_0 \cdot e^{-\Delta t/RC}$ during the time interval Δt is much smaller.

5.8.4 Cascade Circuit

For many special applications, in particular for particle accelerators (see Vol. 4, Chap. 3) one needs very high dc-voltages, which cannot be realized by the rectification circuits discussed so far, because the dielectric strength of transformers sets an upper limit for the secondary voltage.

To overcome this problem *Greinacher* (1880–1974) developed an ingenious circuit based on a cascade of rectifying diodes and capacitors. In Fig. 5.51 such a circuit is shown for the example of 6 diodes and capacitors. Its understanding demands a more thorough consideration:

The lower end S_0 of the secondary coil of the transformer in Fig. 5.51 is grounded. During the negative half cycle in S_1 the voltage change is transferred through the capacitor C_1 to the point P_1. Since the diode D_1 passes the negative voltage in P_1 to S_0 it shortens the voltage between P_1 and S_0 and keeps the voltage in P_1 grounded, while S_1 is at the voltage $-U_0$. During the next half cycle the voltage in S_1 increases from $-U_0$ to $+U_0$. This voltage step of $2U_0$ is transmitted from C_1 to P_1. The voltage in P_1 is now $+2U_0$. It is transferred through D_2, D_3, D_4, D_5, D_6 to the points P_2 – P_6 where at each of these points the voltage $2U_0$ exists. During the next half cycle the voltage in S_1 decreases again to $-U_0$, but in P_1 only to $U = 0$ because of the shortening by D_1. In P_3 the voltage decreases to $+U_0$ because the capacitor C_3 transfers the voltage step $\Delta U = -U_0$ in P_1 completely to P_3.

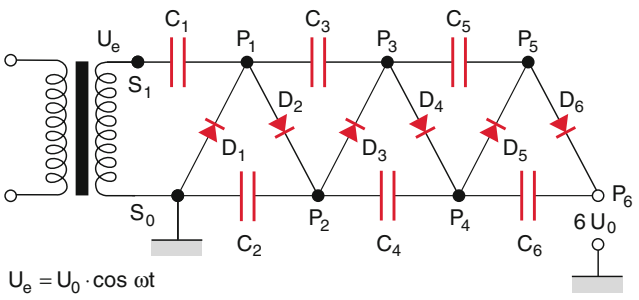


Fig. 5.51 Cascade circuit for the multiplication of the rectified dc-voltage

During the next positive half cycle in S_1 there is again a voltage step of $\Delta U = 2U_0$ in S_1 which is transferred by the diodes to the points P_3 – P_6. Now the voltage in P_1 is $+2U_0$ in P_2 – P_6, however, already $+3U_0$. After each full cycle the voltage increases by $+U_0$ until finally the voltage in the point P_n has increased to $n \cdot U_0$. In our example with n = 6 to $+6U_0$. For more details see [6].

5.9 Electron Tubes

Electron tubes consist of an evacuated glass bulb which contains several electrodes connected by electric penetration to the outside by conductive wires melted into the bottom of the tube.

Although nowadays electron tubes have been mainly replaced by semiconductor devices, they were in former times indispensable tools for the development of modern electronics and they are still in use as high power sources for broadcasting stations. It is therefore worthwhile to study their basic principles.

5.9.1 Vacuum Diodes

The simplest electron tube is the diode, consisting of only two electrodes, the heated cathode and the anode (Fig. 5.52). The heated cathode emits electrons which are accelerated to the anode if it has a positive voltage against the cathode. The electron current through the diode depends on the

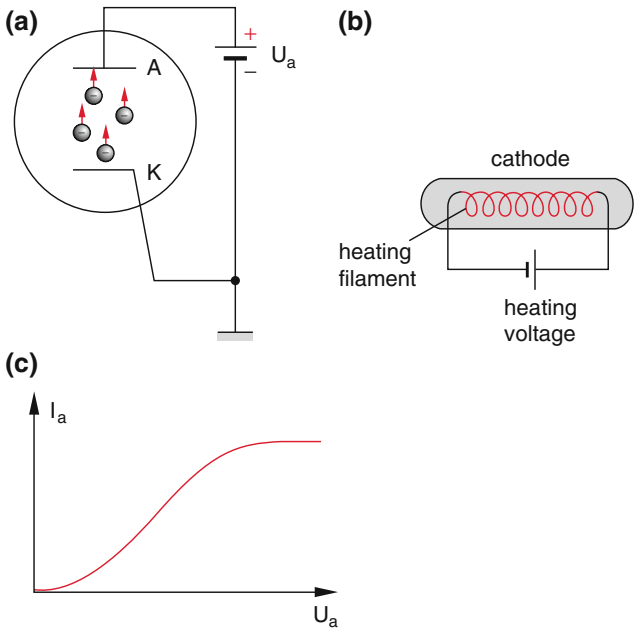


Fig. 5.52 Vacuum diode **a**) circuit, **b**) heated cathode filament, **c**) current-voltage characteristic

temperature of the cathode and the anode voltage U_a. The current increases at first with increasing voltage (Fig. 5.52c) until it reaches a saturation value when all electrons emitted by the cathode are collected on the anode. For negative values of U_a the electrons are repelled by the anode and the current is zero, if the kinetic energy of the electrons is smaller than the repelling potential $-e \cdot U_a$. Vacuum diodes can be therefore used as rectifiers [7].

5.9.2 Triodes

Triodes contain besides cathode and anode a third electrode, the *control grid G*. (Fig. 5.53). The control grid consists of a cylindrical mesh, which encloses the cathode. The electrons must pass on their way from the cathode to the anode through the meshes of the control grid. If the voltage of the control grid U_g is negative against the cathode, the electrons cannot reach the anode and the anode current becomes zero. The control grid can therefore control the anode current by small changes of its voltage (Fig. 5.53b). As long as $U_g < 0$ the anode current can be powerless controlled.

If in addition to the dc-voltage a small ac-voltage

$$U_g = U_{g0} + a \cdot \cos \omega t,$$

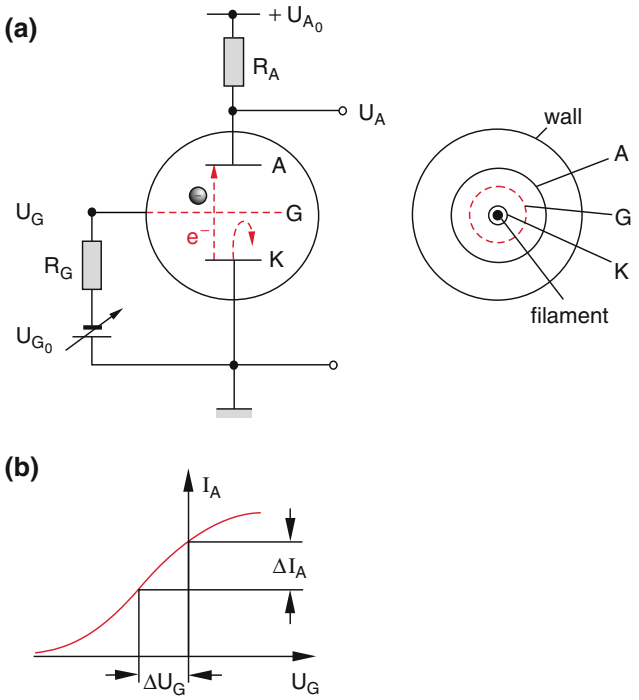


Fig. 5.53 Triode. **a**) Circuit, **b**) influence of grid voltage on anode current

is applied to the control grid, the anode current is modulated according to

$$I_a = I_{a0} + b \cdot \cos \omega t.$$

If the anode is connected to the positive voltage supply with $U = U_{a0}$ through a resistor R_a, the voltage between anode and cathode is

$$U_a = U_{a0} - R_a \cdot I_a.$$

Which is also modulated with the modulation amplitude (Fig. 5.54)

$$\Delta U_a = -R \cdot b \cdot \cos \omega t,$$

The amplitude ΔU_a of the modulated anode voltage is generally much larger than the modulation amplitude $\Delta U_g = a \cdot \cos \omega t$ of the control grid. The voltage amplification

$$V_U = \frac{R_a \cdot b}{a}$$

depends on the operation parameters U_{g0}, U_{a0}, R_a and on the geometrical structure of the triode. One obtains values $V = 10$ up to $V = 1000$. Figure 5.55 shows a realistic drawing of a triode [8] and Fig. 5.56 a schematic drawing of a tetrode.

The tetrode has an additional grid and Fig. 5.56 gives a schematic drawing of a tetrode (screening grid), which has a positive voltage against the cathode. It shields the control grid against the anode. Above a threshold value the anode

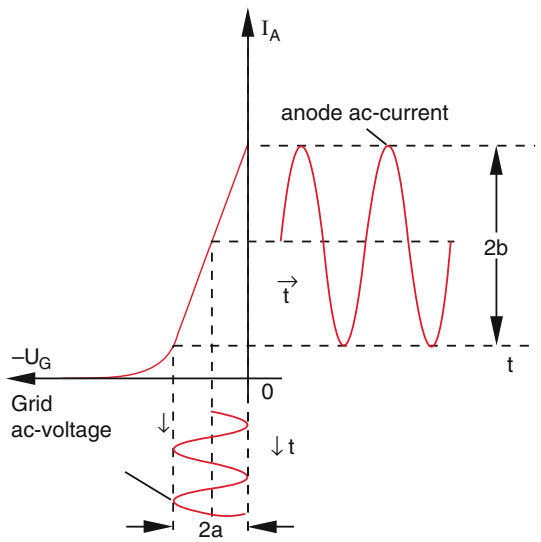


Fig. 5.54 Modulation of anode current by the modulated grid voltage

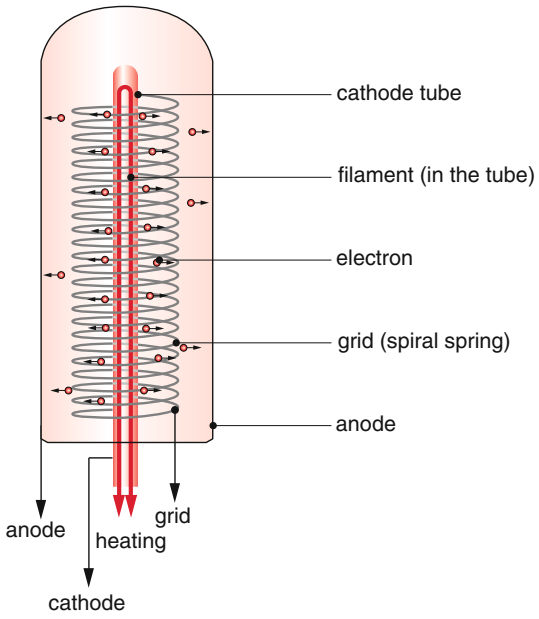


Fig. 5.55 Real design of a triode

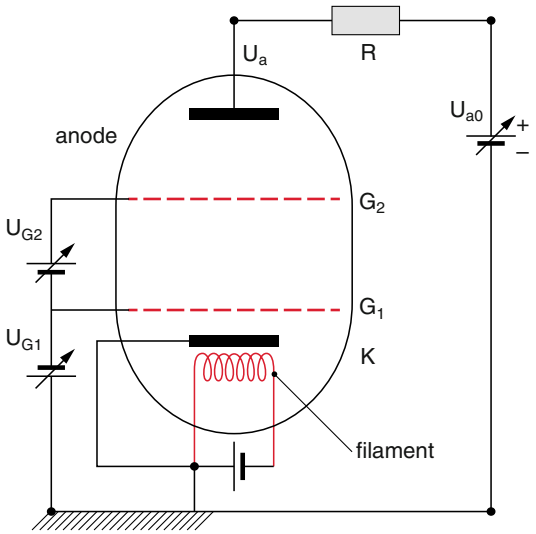


Fig. 5.56 Tetrode. G_1 = conrol grid, G_2 = screen grid

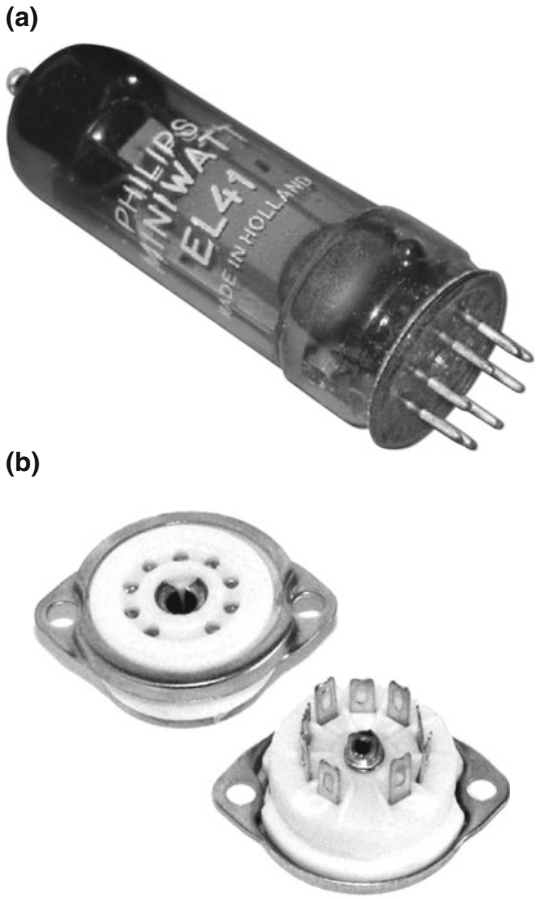


Fig. 5.57 **a**) Commercial pentode with tube socket, **b**) socket of an electron tube. Upper part: side towards the tube; lower part: side of the circuit

current is nearly independent of the anode voltage and is essentially determined by the voltage of the control grid. Tetrodes are used for high power applications in high frequency technology (Fig. 5.56).

The connections to the different electrodes are realized by metal pins, which fit into standardized sockets (Fig. 5.57). From here the connections run to the different components of an electric circuit [9].

Summary

- The mechanical torques excerted onto coils with electrical current in a magnetic field are used in electromotors to perform mechanical work.
- Electrical generators generate an ac-induction voltage by turning a coil in a magnetic field.
- The average power of an ac-current is

$$\bar{P} = U_{\text{eff}} \cdot I_{\text{eff}} \cdot \cos\varphi = \tfrac{1}{2} U_0 \cdot I_0 \cdot \cos\varphi,$$

- where φ is the phase shift between current and voltage.
- A three-phase ac- current generates a rotating magnetic field which is used for the electric drive of electro-motors.
- An electric ac-resonant circuit consists of resistor R, inductance L and capacitor C. Its resonance frequency is

$$\omega_{\text{r}} = 1/\sqrt{L \cdot C}.$$

- Series and parallel resonant circuits show with respect to their complex resistance Z a complimentary behavior: At the resonance frequency ω_{r} is Z real and maximum for parallel but minimum for series circuits.
- Transformers consist of two solenoids coupled by an iron core. They transform the input voltage U_{e} into an output voltage U_{a} with the ratio $U_{\text{a}}/U_{\text{e}} = N_2/N_1$ which equals the ratio of the number of windings N_2 in the output coil and N_1 in the input coil.
- AC-voltages are rectified by diodes. The commonly used circuit with 4 diodes in a bridge arrangement is called *Graetz*- circuit. The output dc-voltage must be smoothed by using L-C voltage dividers or an electronic stabilization device.
- In an appropriate arrangement of capacitors and rectifiers the output voltage can be a multiple of the input voltage.
- Input ac-voltages and currents can be amplified by transistors or electron tubes (triodes) in special circuits.

Problems

5.1 An electronic circuit in a black box consists of a resistor R and a capacitor C and input and output sockets (Fig. 5.58). For a dc-input voltage it has a resistance of 100 Ω, for an ac-voltage with 50 Hz its resistance is 20 Ω. Determine the circuit and the values of R and C.

5.1b A frequency filter in the black box of Fig. 5.58 has its maximum transmission $T = |U_2|/|U_1|$ at the frequency $\omega = 75\ s^{-1}$. The transmission at $\omega = 0$ is $T = 0.01$. It consists of resistor R, inductance $L = 0.1$ H with $R_L = 1\ \Omega$ and capacitor C. How is the circuit set up and what are the values of R and C?

5.2 Calculate the frequency dependent complex resistance $Z(\omega)$ and its amount $|Z(\omega)|$ for the parallel electrical resonant circuit in Fig. 5.35a. What are the resonance frequency ω_0 ad frequency halfwidth $\Delta\omega$ with $R_L = 1\ \Omega$, L = 10^{-4} H and $C = 1\ \mu F$?

5.3 A transformer without iron core consists of two long solenoids with N_1 and N_2 windings with cross section area A; which are tightly wound about one another. Determine the secondary voltage U_2, current I_2 and the phase shift $\Delta\varphi$ against the primary voltage U_1 when the transformer output is loaded

(a) with a resistor R
(b) with a capacitor C

What is the input power, if the losses in the transformer are negligible?

5.4 Calculate for the circuit in Fig. 5.59 the transmission $|U_2|/|U_1|$ of the voltage and $|I_2|/|I_1|$ of the current at the input voltage $U_1 = U_0 \cdot \cos \omega t$ with $\omega = 300\ s^{-1}$, $L = 0.1$ H, $C = 100\ \mu F$ and $R = 50\ \Omega$.

5.5 A flat circular coil with $N = 500$ windings and an area $A = 100\ cm^2$ rotates in a homogeneous magnetic field $B = 0.2$ mT about an axis in the coil plane (Fig. 5.1). What is the mechanical power one has to apply for rotating the coils at a frequency f = 50 Hz to deliver the electrical output into the resistor $R = 10\ \Omega$, if the coil resistance is $R_L = 5\ \Omega$?

5.6 An ac-source $U = U_0 \cdot \cos \omega t$ with $\omega = 2\pi \cdot 50\ s^{-1}$ and $U_0 = 15$ V is connected to

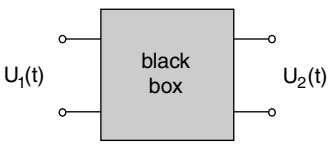


Fig. 5.58 Illustration of problem 5.1 b

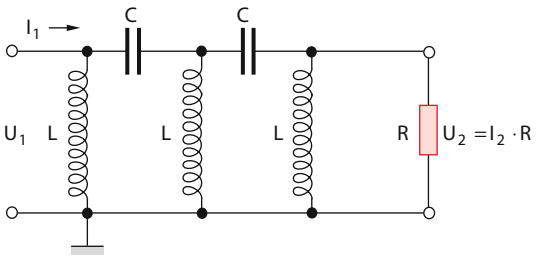


Fig. 5.59 Illustration of problem 5.4

(a) a one-way rectifying circuit (Fig. 5.45)
(b) a Graetz rectifying circuit (Fig. 5.48).
Illustrate for a load resistor $R = 50\ \Omega$ and a capacitor $C = 1$ mF the temporal course of the output voltage $U_2(t)$ with the residual ripples and the dc load power.

5.7 A capacitor $C = 10\ \mu F$ with a leak resistance of 10 MΩ is connected to an ac-voltage source $U = U_0 \cdot \cos \omega t$ with $U_0 = 300$ V and $\omega = 2\pi \cdot 50\ s^{-1}$. What is the total current (real- plus blind current) and which power is consumed in the capacitor?

5.8 The ac-circuit in Fig. 5.29 is connected to the source $U = U_0 \cdot \sin \omega t$. What is the voltage U_L across the inductance L (amplitude and phase)?
Numerical example: $R = 20\ \Omega$, $L = 0.05$ H, $C = 50\ \mu F$, $U_0 = 300$ V and $\omega = 2\pi \cdot 50\ s^{-1}$.

5.9 Calculate the transmission (U_a/U_e) if the L-C filtering element of Fig. 5.49 is replaced by the low pass filter in Fig. 5.33.

5.10 Derive Eq. (5.7) and calculate the load current I_2 which gives the maximum terminal voltage U_K of the shut-wound generator.

References

1. https://en.wikipedia.org/wiki/Electric_generato
2. C.Rawlins: Basic AC-Circuits (Newnes 2000)
3. St. Winder: Analog and Digital Filter Design, (Newnes 2002)
4. J.H.Harlow: Electric Power Transformers (CRC Press 2012)
5. https://en.wikipedia.org/wiki/Rectifier
6. https://en.wikipedia.org/wiki/Voltage_multiplier
7. https://en.wikipedia.org/wiki/Vacuum_tube#Diodes
8. Sogo Okamura (ed), *History of Electron Tubes*, IOS Press, 1994 ISBN 90-5199-145-2 page 20
9. B. Rosenblit: Tubes and Circuits (Create Space independent Publishing Platform 2012

6 Electromagnetic Oscillations and the Origin of Electromagnetic Waves

The next two chapters are very important, not only for electro-technical applications but even more for the basic understanding of the generation and the propagation of electro-magnetic waves. The mathematical treatment is quite similar to that of mechanical waves, which has been extensively discussed in Vol. 1, Chap. 11.

6.1 The Electromagnetic Oscillating Circuit

An electromagnetic oscillating circuit consists of a capacitor C, an inductance L and an Ohmic resistor R (see Sect. 5.4), where the capacitor is periodically charged and discharged. The comparison with a mechanical oscillating circuit is illustrated in Fig. 6.1 for the model of an oscillating mass m, that is bound by spring-forces to its equilibrium location (harmonic oscillator Vol. 1, Sect. 11.1).

The maximum potential energy of the mass m corresponds to the electrical energy $W_{\mathrm{el}} = 1/2C \cdot U^2$ of the charged capacitor (Fig. 6.1a). The capacitor discharges through the inductance L and the resulting current $I = \mathrm{d}Q/\mathrm{d}t$ generates in the inductive coil L a magnetic field B with the magnetic energy $W_{\mathrm{m}} = 1/2L \cdot I^2$, which corresponds to the kinetic energy $1/2m \cdot v^2$ in the mechanical model. Because of its inertial mass the mass m moves through the equilibrium point to the other side and transfers its kinetic energy again into potential energy. For the electrical circuit the induction law and Lenz's Rule are the analogue to the inertia. When the current I decreases, an induction voltage is generated in the coil which hinders the decrease of the current I. The current I is driven by the induction voltage into the capacitor C until C is completely charged again (Fig. 6.1c). Now the procedure is repeated in the opposite direction.

6.1.1 Damped Electromagnetic Oscillations

Analogue to the mechanical model where the friction causes the damping of the oscillation, in the electromagnetic circuit any Ohmic resistance R of the coil or the windings cause a loss of the electric energy. The decrease of the electric energy per sec is $\mathrm{d}W_{el}/\mathrm{d}t = I^2 \cdot R$ which is converted into heat energy. The result is a damped electromagnetic oscillation (Fig. 6.2).

We regard as example the series circuit in Fig. 5.29. When the circuit is excited to oscillations by an external pulse (Fig. 6.2a) it performs after the end of the pulse ($U_{\mathrm{e}} = 0$) damped electro-magnetic oscillations. Their mathematical treatment starts from Eq. (5.21).

$$L \cdot \frac{\mathrm{d}^2 I}{\mathrm{d}t^2} + R \cdot \frac{\mathrm{d}I}{\mathrm{d}t} + \frac{1}{C} I = 0. \tag{6.1}$$

We try (completely analogue to the mechanical treatment in Vol. 1, Sect. 11.4) the ansatz

$$I = A \cdot \mathrm{e}^{\lambda t}\,, \tag{6.2a}$$

where A and λ can be complex quantities. Inserting into (6.1) gives the equation for λ

$$\lambda^2 + \frac{R}{L}\lambda + \frac{1}{LC} = 0\,, \tag{6.3a}$$

with the solutions

$$\begin{aligned} \lambda_{1,2} &= -\frac{R}{2L} \pm \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}} \\ &= -\alpha \pm \beta \end{aligned} \tag{6.3b}$$

which depends on the value of α, i.e. on the ratio R/L.

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_6

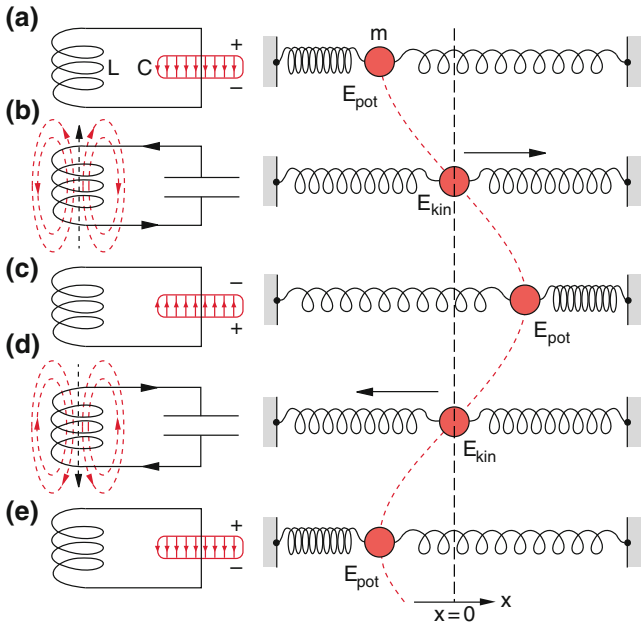


Fig. 6.1 Comparison between electro-magnetic oscillation circuit and the mechanical model of an oscillating mass suspended between two springs

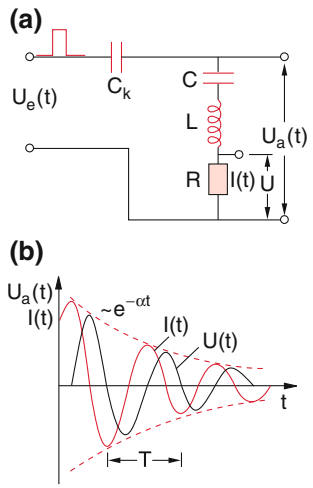


Fig. 6.2 Damped oscillating circuit, excited at $t = 0$ by an electric pulse coupled to the circuit through a capacitance. **a**) Experimental realization for the measurement of current I and voltage U. **b**) Temporal dependence of $U_a(t)$ and $I(t) = U(t)/R$

The general solution of (6.1) is

$$I = A_1 e^{-(\alpha-\beta)t} + A_2 e^{-(\alpha+\beta)t} . \tag{6.2b}$$

We will now discuss some cases with special initial conditions.

6.1.1.1 Overdamped Case

For $R^2/(4L^2) > 1/(LC)$ the number β is a real quantity. Since the current I as a real physical quantity must be real it follows that A_1 and A_2 both have to be real. With the initial conditions $I(0) = I_0$ and $dI/dt(0) = dI_0/dt$ we obtain from (6.2a)

$$\begin{aligned} A_1 &= \frac{I_0}{2}\left(1+\frac{\alpha}{\beta}\right) + \frac{\dot{I}_0}{2\beta}, \\ A_2 &= \frac{I_0}{2}\left(1-\frac{\alpha}{\beta}\right) + \frac{\dot{I}_0}{2\beta}. \end{aligned} \tag{6.4a}$$

For the initial condition $dI_0/dt = 0$ the special solution is obtained

$$I(t) = I_0 \cdot e^{-\alpha t}\left[\cosh(\beta t) + \frac{\alpha}{\beta}\sinh(\beta t)\right]. \tag{6.4b}$$

The current drops monotonically from $I = I_0$ to $I = 0$ which is reached only for $t = \infty$ (curve (a) in Fig. 6.3). For the case $I_0 = 0$ but $dI_0/dt \neq 0$ the solution is

$$I(t) = (\dot{I}_0/\beta) \cdot e^{-\alpha t} \sinh(\beta t) . \tag{6.4c}$$

The current increases initially from $I(0) = 0$ to a maximum and crepes then asymptotically towards $I = 0$ (curve b in Fig. 6.3).

6.1.1.2 Aperiodic Limiting Case

For $\beta = 0$ we have the aperiodic limiting case with the solution (see Vol. 1, Chap. 11)

$$I(t) = e^{-\alpha t}(I_0 + A_3 t) \tag{6.5a}$$

With the constant $A_3 = \alpha \cdot I_0 + dI_0/dt$.
For $I_0 = 0$ Eq. (6.5a) reduces to

$$I(t) = \dot{I}_0 \cdot t \cdot e^{-\alpha t} \tag{6.5b}$$

(red curve c in Fig. 6.3). With these initial conditions $I(t)$ does not cross the line $I = 0$ but reaches $I = 0$ only asymptotically.

For another initial condition $(I(0) \neq 0, dI_0/dt(0) \neq 0$ $I(t)$ cuts the t-axis $I = 0$ for finite values of t and approaches again zero for $t = \infty$ (curve d in Fig. 6.3.

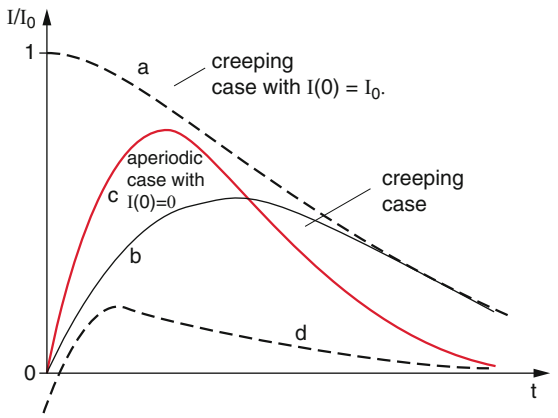


Fig. 6.3 Limiting cases of the damped oscillation. **a**) Creeping case with $I(0) = I_0 \neq 0$. **b**) Creeping case with $I(0) = 0$, **c**) aperiodic limiting case, **d**) aperiodic case with $I(0) < 0$

6.1.1.3 Damped Oscillation

The most important case is realized for $R^2 < 4L/C$, where β becomes imaginary. We set $\beta = i \cdot \omega$ and obtain with (6.3a, 6.3b) the solution of (6.1) as

$$I(t) = e^{-\alpha t}[A_1 e^{i\omega t} + A_2 e^{-i\omega t}], \tag{6.6}$$

where the coefficients $A_1 = a + i \cdot b$ and $A_2 = a - i \cdot b$ are complex conjugates of each other. Therefore the current $I(t)$ becomes a real physical quantity. Inserting the expressions for A_i converts (6.6) to

$$I(t) = 2|A| \cdot e^{-\alpha t} \cos(\omega t + \varphi) \tag{6.7}$$

With $|A| = \sqrt{(a^2 + b^2)}$ and $\tan \varphi = b/a$. The values of a and b are determined by the initial conditions.

The current $I(t)$ in the circuit performs a damped oscillation with the resonance frequency

$$\omega_R = \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}, \tag{6.8a}$$

which becomes for $R = 0$ the frequency $\omega_0 = 1/\sqrt{L \cdot C}$ of the undamped circuit. The oscillation period of the circuit is

$$T = \frac{2\pi}{\omega_R}. \tag{6.8b}$$

Example

$L = 10^{-2}$H; $C = 10^{-6}$F; $R = 100\,\Omega \Rightarrow \omega = 8.6 \times 10^3$ rad/s $\Rightarrow \nu = 1.4$kHz $\Rightarrow T = 0.7$ ms. With $R = 0$ these values would slightly change to $\nu_0 = 1.6$kHz; $T_0 = 0.63$ms.

6.1.2 Forced Oscillations

When the series circuit in Fig. 6.4a is connected to an external ac-voltage $U(t) = U_0 \cdot \cos \omega t$ the circuit oscillates with the stationary amplitude U_0 and also the current $I(t) = I_0 \cdot \cos(\omega t - \varphi)$ has a temporally constant amplitude $I_0 = U_0/|Z|$, where Z is the complex resistance of the circuit introduced in Sect. 5.4. The real electric power converted in the resistor R into heat is

$$P_{el}^{real} = I^2 R = \frac{U^2}{Z^2} \cdot R = \frac{[U_0 \cdot \cos(\omega t)]^2}{Z^2} \cdot R. \tag{6.9}$$

Inserting for Z the expression (5.25) we obtain with $\langle \cos^2 \omega t \rangle = 1/2$ the average loss of electric power

$$\langle P_{el}^{real} \rangle = \frac{1}{2} \cdot \frac{U_0^2 \cdot R}{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2}. \tag{6.10}$$

The power loss of the oscillating series circuit reaches its maximum

$$\langle P_{el}^{real} \rangle_{max} = \frac{1}{2} \frac{U_0^2}{R}. \tag{6.11}$$

for the resonance frequency $\omega = \omega_0 = \omega_R = 1/\sqrt{L \cdot C}$ of the undamped circuit. The resistance $Z(\omega_0) = R$ reaches its minimum.

In Fig. 6.4a the frequency-dependent power loss is plotted for the weakly damped series circuit. The full half width of the resonance curve $\Delta P_{el}(\omega)$ (this is the frequency difference $\Delta\omega = \omega_1 - \omega_2$ between the half points $I(\omega_1) = I(\omega_2) = 1/2\, I(\omega_0)$) is for $R/\omega L \ll 1$: $\Delta\omega_{1,2} \approx R/L$.

Similar conditions are obtained for the parallel circuit of Fig. 6.4b. However, here the resistance $Z(\omega_0) = R$ takes its *maximum* value at $\omega = \omega_0$ and the power loss its minimum (see problems 5.8 and 6.2a, 6.2b).

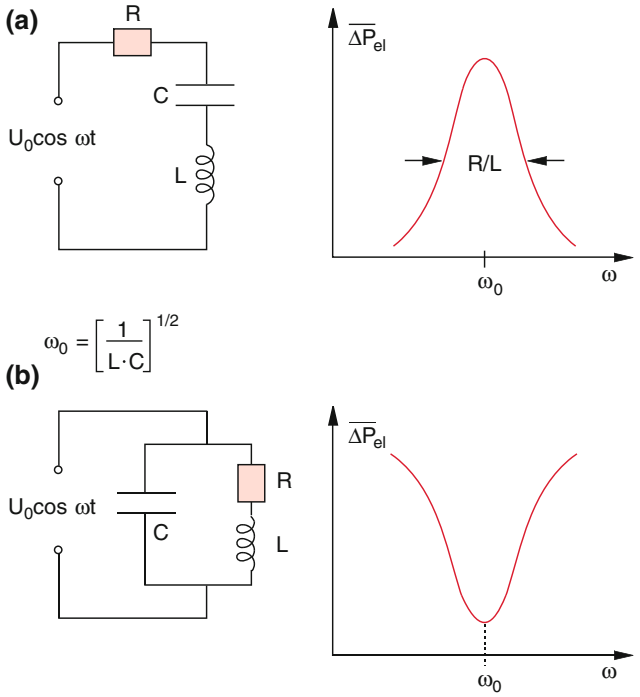


Fig. 6.4 Frequency-dependent energy loss in oscillating circuits connected to a source with periodic voltage. **a**) series circuit, **b**) parallel circuit

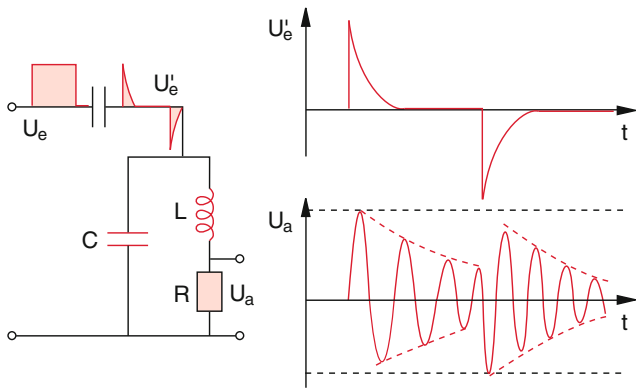


Fig. 6.5 Excitation of a damped parallel oscillation circuit by a sequence of electric pulses

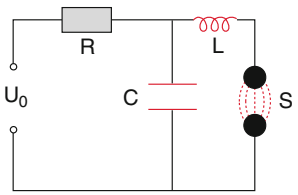


Fig. 6.6 Spark oscillation circuit

An experimental example for the excitation of damped oscillations is illustrated in Fig. 6.5. The circuit is excited by a regular sequence of electric pulses. The damping can be controlled by the choice of the ratio R/L. With this arrangement the special cases of damped and over damped oscillations or the aperiodic limiting case can be readily demonstrated on the oscilloscope screen, just by selecting the proper ratio R/L.

Historically the first generation of damped electrical oscillations was realized by the spark oscillation circuit shown in Fig. 6.6. The capacitor C is charged by a dc-source with voltage U_0 through the resistor R. As soon as the voltage U at the capacitor exceeds the ignition voltage U_c of a spark gap S the gap ignites. The discharge current $I = \mathrm{d}Q/\mathrm{d}t$ generates in the inductance coil a magnetic field, which drives during the decrease of the field the current for recharging the capacitor C again (as shown in Fig. 6.1). This leads to a damped oscillation in the C, L, S circuit. The damping is due to the resistance R of the ignited spark gap. If the oscillation frequency is sufficiently high the electric conductivity of the spark gap is preserved even at the zero crossing of the current $I(t)$, because the ions in the spark do not recombine fast enough nor do they leave the spark during half an oscillation period. The conductivity therefore never becomes zero.

6.2 Coupled Oscillation Circuits

Analogue to mechanical oscillators coupled by springs (see coupled pendula in Sect. 11.8 of Vol. 1) also electrical oscillation circuits can be coupled either by inductive coupling, capacitive coupling or by an Ohmic resistor. This coupling causes a partial transfer of the oscillation energy from one circuit to the second one and back.

As example Fig. 6.7 shows the inductive coupling of two electrical oscillation circuits. In addition to the induction voltage $U_{\mathrm{ind}} = -\mathrm{d}I/\mathrm{d}t$ of the uncoupled circuit now a voltage $U_1 = -L_{12}\mathrm{d}I_2/\mathrm{d}t$ has to be added in the first circuit and $U_2 = -L_{12}\mathrm{d}I_1/\mathrm{d}t$ in the second circuit. Instead of the Eq. (5.21) we now obtain the coupled differential equations

$$L_1\frac{\mathrm{d}^2I_1}{\mathrm{d}t^2} + R_1\frac{\mathrm{d}I_1}{\mathrm{d}t} + \frac{I_1}{C_1} = -L_{12}\frac{\mathrm{d}^2I_2}{\mathrm{d}t^2} \tag{6.12a}$$

$$L_2\frac{\mathrm{d}^2I_2}{\mathrm{d}t^2} + R_2\frac{\mathrm{d}I_2}{\mathrm{d}t} + \frac{I_2}{C_2} = -L_{12}\frac{\mathrm{d}^2I_1}{\mathrm{d}t^2} \tag{6.12b}$$

Inserting $I_k = I_{0,k}\cdot \mathrm{e}^{\mathrm{i}\omega t}$ (k = 1, 2) we get the two coupled equations

$$\begin{aligned}\left(-L_1\omega^2 + \mathrm{i}\omega R_1 + \frac{1}{C_1}\right)I_1 - \omega^2 L_{12}I_2 &= 0\\ -\omega^2 L_{12}I_1 + \left(-L_2\omega^2 + \mathrm{i}\omega R_2 + \frac{1}{C_2}\right)I_2 &= 0\end{aligned} \tag{6.13}$$

for the currents $I_1(t)$ and $I_2(t)$. The equations have nontrivial solutions only if the determinant of the coefficients is zero. This gives the equation for the determination of the resonance frequency of the two coupled circuits

$$\left[R_1 + \mathrm{i}\left(\omega L_1 - \frac{1}{\omega C_1}\right)\right]\cdot\left[R_2 + \mathrm{i}\left(\omega L_2 - \frac{1}{\omega C_2}\right)\right] = \omega^2 L_{12}^2, \tag{6.14}$$

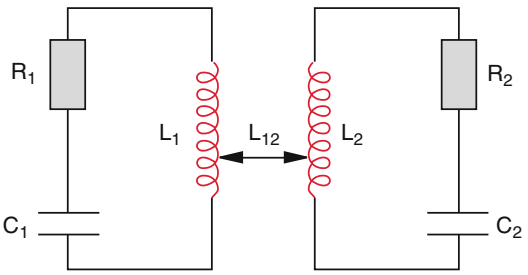


Fig. 6.7 Inductively coupled oscillating circuits

The general solution of this equation is a bit tedious. We will therefore only discuss the more simple special case of two coupled equal circuits with no losses ($R_1 = R_2 = 0$; $L_1 = L_2 = L$; $C_1 = C_2 = C$). With the degree of coupling $k = L_{12}/L$ (see Sect. 5.6) we obtain from (6.14) as solution of the quadratic equation for ω^2 the two frequencies

$$\omega_1 = \sqrt{\frac{1}{(L - L_{12})C}} = \frac{\omega_0}{\sqrt{1 - L_{12}/L}} = \frac{\omega_0}{\sqrt{1-k}}, \tag{6.15a}$$

$$\omega_2 = \sqrt{\frac{1}{(L + L_{12})C}} = \frac{\omega_0}{\sqrt{1+k}} \; . \tag{6.15b}$$

The coupling causes a splitting of the resonance frequency ω_0 of the uncoupled circuits into the two frequencies ω_1 and ω_2. For weak coupling ($k \ll 1$) the frequency difference becomes

$$\Delta\omega = \omega_1 - \omega_2 = \omega_0 \cdot k = \omega_0 \cdot \frac{L_{12}}{L}, \tag{6.16}$$

It is proportional to the degree k of coupling. Compare the completely analogue conditions for mechanical oscillators (Vol. 1, Sect. 11.8).

Besides the inductive coupling also the capacitive coupling by a common capacitor is used (Fig. 6.8a) or the galvanic coupling through a common resistor R (Fig. 6.8b). Their mathematical treatment is similar to that of inductive coupling. Instead of the coupling term $\omega^2 L_{12}$ now the terms $1/\omega C$ for capacitive or $\omega \cdot R$ for galvanic coupling describe the coupling [1].

If the first of the two inductively coupled circuits in Fig. 6.7 with the complex resistances $Z_i = R_i + i \cdot (\omega L_i - 1/\omega C_i)$ (i = 1, 2) is connected to the ac-voltage $U = U_0 \cdot e^{i\omega t}$ one obtains instead of (6.13) after division by $i\omega$ the equations

$$\begin{aligned} U &= Z_1 I_1 + i\omega L_{12} I_2 \; , \\ 0 &= i\omega L_{12} I_1 + Z_2 I_2 \; . \end{aligned} \tag{6.17}$$

Elimination of I_1 gives the current I_2 in the second circuit

$$I_2 = -\frac{i\omega L_{12}}{\omega^2 L_{12}^2 + Z_1 Z_2} U \; . \tag{6.18}$$

Inserting the expressions for Z_i one obtains a rather long expression. With the abbreviations $X = \mathrm{Im}(Z) = \omega L - 1/\omega C$ it can be simplified for coupled equal circuits ($Z_1 = Z_2 = Z$) and gives

$$|I_2| = \frac{\omega L_{12}}{\sqrt{[\omega^2 L_{12}^2 + R^2 - X^2] + 4R^2X^2}} |U|. \tag{6.19a}$$

Remark In Electro-technics the symbol X is used for the *reactance* Im(Z).

For lossless circuits (R = 0) we obtain from (6.19a) the ratio

$$\frac{|I_2|}{|U|} = \frac{\omega^3 k/L}{\omega^4(k^2 - 1) + 2\omega_0^2\omega^2 - \omega_0^4}, \tag{6.19b}$$

where $\omega_0 = 1/\sqrt{L \cdot C}$ is the resonance frequency of the uncoupled circuit. In Fig. 6.9 the ratio $|I_2|/|U|$ is plotted as a

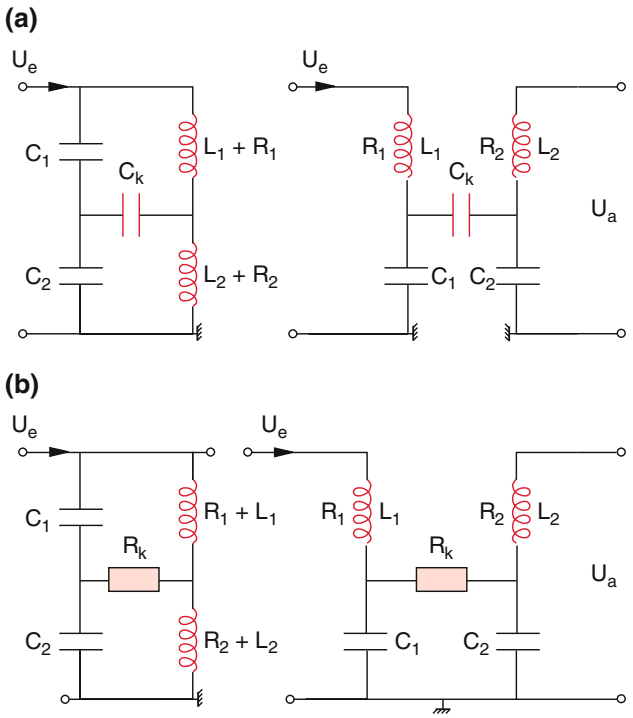


Fig. 6.8 **a**) Capacitive coupling of parallel and series oscillating circuits, **b**) galvanic coupling of oscillating circuits

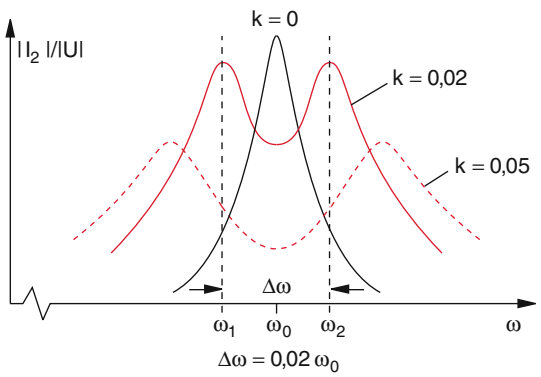


Fig. 6.9 Resonance curve for the current $I_2(\omega)$ in a parallel oscillating circuit which is coupled to another circuit driven by an external voltage $U = U_0 \cdot \cos \omega t$

function of the frequency ω for different degrees of coupling. The figure shows that for $k \neq 0$ two resonance frequencies are present with a distance $\Delta\omega$ that increases with increasing degrees of coupling k.

6.3 Generation of Undamped Oscillations

In order to realize undamped oscillations even in circuits with losses the energy loss of the circuit has to be compensated by external energy supply. This can be realized in different ways.

A simple examples which can be realized only for very slow oscillations but is very impressive for demonstrations, uses the manual action on a switch, which supplies the missing energy from an external source to the capacitor always at the correct time. Inductance L and capacitor C are chosen so large that the oscillation frequency is about 1 Hz. The phase-shifted oscillating current and voltage can be then demonstrated on two large meters which can be viewed even by a large auditorium. Using a light bulb as resistor the periodic oscillations of its brightness visualize the periodic oscillations of the current in the circuit (Fig. 6.10).

For higher frequencies the reaction speed of the human brain is too low and electronic devices have to be used. One example is the *Meißner*-circuit shown in Fig. 6.11. Here the dc-current supply is connected to the circuit by inductive coupling between the two coils L and L_f, where L_f generates the feedback between grid G of the triode and the cathode. If the grid voltage U_G is negative against the cathode voltage the electrons, emitted by the cathode cannot reach the anode [2].

Any small perturbation can induce a small oscillation in the circuit, which is transferred to the grid by the inductive coupling. It changes periodically the voltage of the grid and generates a periodically changing anode current, which produces in the coil L a modulated magnetic field and a modulated voltage that is transferred to the grid. The oscillation

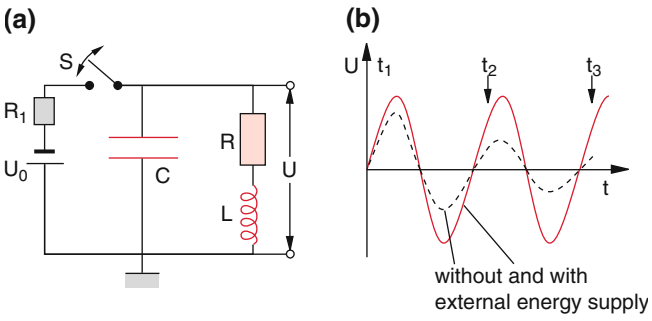


Fig. 6.10 Realization of slow undamped oscillations of a damped oscillation circuit by manual closing the switch S periodically at times $t = t_0 + n \cdot \Delta t$ with $\Delta t = T$ (oscillation period). **a**) Experimental arrangement, **b**) oscillations with and without periodic energy supply

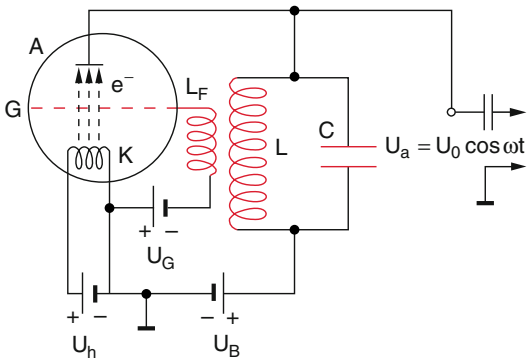


Fig. 6.11 Meißner feedback design for the generation of undamped oscillations in the radio frequency range

amplitude increases until a stable oscillation $U = U_0 \cos \omega t$ of the circuit is reached which depends on the voltage U_a and the grid bias voltage U_G.

The undamped oscillations are not restricted to cosωt or sinωt oscillations but can have any time dependence. As example Fig. 6.12 shows a circuit that generates a periodic saw tooth voltage. When the switch S is closed at time $t = 0$ the source with the dc-voltage U_0 charges the capacitor C until the ignition voltage U_Z of the glow discharge G is reached where the glow discharge lamp G ignites. Since the resistance of the ignited discharge lamp is very small compared to the resistor R in the charging line ($R_G \ll R$) the capacitor C discharges and its voltage drops until the extinction voltage is reach, where the discharge extinguishes. Now the charging process starts again. From Eq. (2.11) one obtains the period T of the saw tooth voltage

$$T = RC \cdot \ln \frac{U_0 - U_L}{U_0 - U_Z}. \tag{6.20}$$

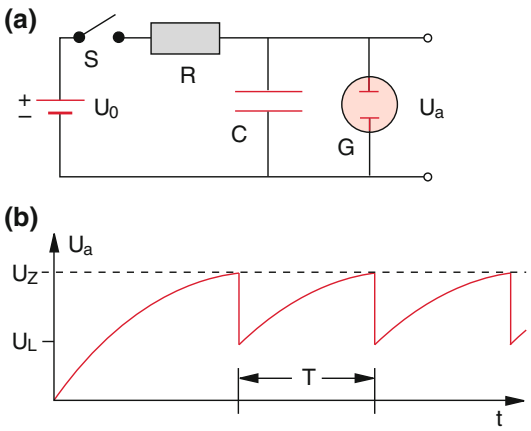


Fig. 6.12 Saw tooth oscillation **a**) experimental design, **b**) time variation of the voltage across the glow discharge lamp

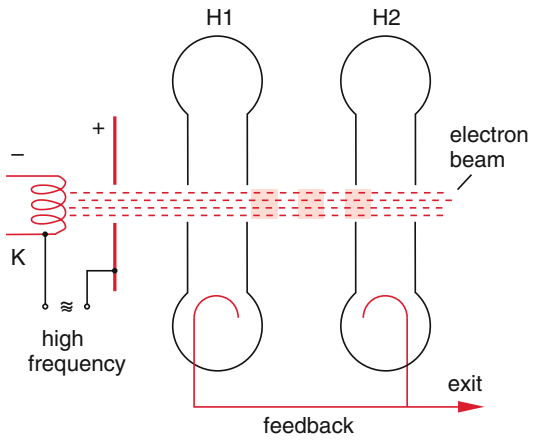


Fig. 6.13 Schematic presentation of the operation of a klystron

For very high frequencies, the capacitance C and the inductance L in the oscillating circuit of Fig. 6.11 are too large. Furthermore electron tubes are no longer capable to induce oscillations with frequencies $\omega > 10^{10}/\text{s}$, because the time of flight of the electrons through the tube from cathode to anode becomes already larger than the period T of the oscillation. Therefore, a new type of tubes has been developed for very high frequencies, called **klystrons** (Fig. 6.13). They consist of two cavity resonators. Their principle can be understood as follows [2]:

When electrons, emitted by the hot cathode are accelerated by a positive dc-voltage U before they enter the first cavity, they reach a velocity $v = \sqrt{(2e \cdot U/m)}$. A high frequency voltage between cathode and anode modulates the velocity of the electrons while they pass through the first cavity. This causes a modulation of the current density $j = \rho \cdot v$ of the electrons when they enter the second cavity. The electrons arrive as periodic charge packets that induce oscillations of the second cavity. This high frequency voltage is coupled with the proper phase back to the first cavity thus amplifying the modulation amplitude of the electron current. This feedback causes the development of a stable oscillation with the resonant frequency of the cavity, starting from random fluctuations of the electron current density in the second cavity. Since such random fluctuations always occur with a broad frequency range, they also contain the resonant frequency of the cavity, which acts as the starting element for the processes mentioned above. Therefore no external HF-voltage is needed to start the oscillations. With the proper choice of the dimensions of the cavities frequencies in the gigahertz range ($\omega = 10^9 – 10^{-12}\ \text{s}^{-1}$) can be realized.

6.4 Open Oscillating Circuits; Hertzian Dipole

In the previous sections we have discussed electromagnetic oscillating circuits, where the energy W_{el} oscillates periodically between electric field energy in capacitors and magnetic field energy in solenoids. Now we will discuss the transition from the closed circuit in Figs. 6.1 and 6.14a, where L and C are still spatially separated, to the open circuit, depicted in Fig. 6.14d. The inductance L of the solenoid in Fig. 6.14a transforms to the inductance of the single conductor loop in Fig. 6.14b. The capacitance C of Fig. 6.14a becomes smaller and smaller when the loop in Fig. 6.14b is bent into the straight wire of Fig. 6.14c. Finally, the end plates in Fig. 6.14c can be completely removed and we are left with a straight wire, which has a small capacitance and inductance.

The essential difference between the closed circuit in Fig, 6.14a and the open circuit in Fig. 6.14d where charges oscillate back and forth through the wire is illustrated in Fig. 6.15. In the closed circuit, the electric and the magnetic field are spatially confined and separated from each other. The main part of the electric field is concentrated between the plates of the capacitor, while the main part of the magnetic field is found inside the solenoid (see Sect. 3.2.6.4). The stray fields outside the capacitor or the solenoid are very weak and can be neglected.

In Fig. 6.15b the electric field is still localized inside the capacitor, but the magnetic field reaches already far out into the whole space. In case of the straight wire in Fig. 6.15c, which carries a high frequency current, the magnetic as well as the electric field reach far out into the whole space. A temporal change of current and charge density in the straight wire also changes the electric and magnetic fields in the surrounding space. These changes propagate with the speed of light into the surrounding space as electromagnetic waves and cause the emission of energy from the wire into space. The arrangement of Fig. 6.15c is called *transmitter*, because it transmits energy, supplied to the straight wire by an external source, into space in form of electromagnetic waves. The straight wire is the *antenna*.

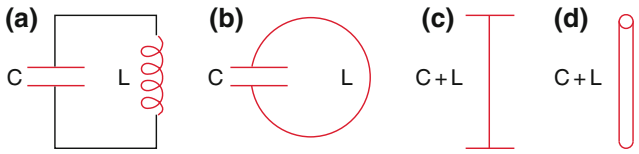


Fig. 6.14 Schematic illustration of the continuous transition from a closed oscillating circuit to the straight wire of an antenna emitting electromagnetic waves into space

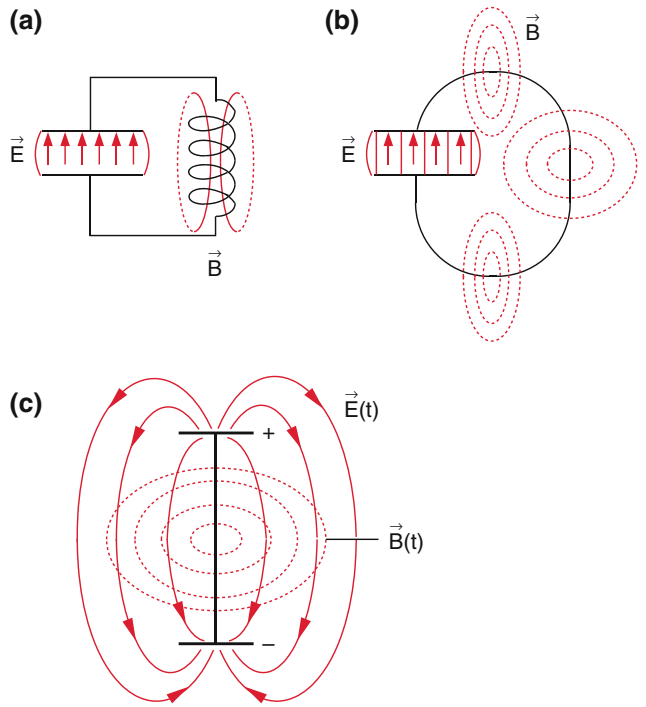


Fig. 6.15 Illustration of the change of the electro-magnetic field during the transition from a closed circuit to the straight wire antenna

We must now answer four questions:

- How can one experimentally realize that electric charges oscillate in a straight wire?
- What is the form of the electric and magnetic fields generated by such an oscillating charge distribution?
- What is the relation between temporally oscillating fields and the electromagnetic waves propagating into space?
- Which radiation power is emitted by the transmitter?

6.4.1 Experimental Realization of a Transmitter

The excitation of electromagnetic oscillations in an open circuit can be realized by inductive, capacitive or galvanic coupling to a closed oscillating circuit that gets its energy loss replaced by an external source. A schematic circuit is shown in Fig. 6.16 and the practical realization in Fig. 6.17. Here the high frequency source is a closed oscillation circuit, where the grid of the triode is coupled by a capacitor to the anode voltage supply, which compensates the energy losses caused by Joule's heat and the energy coupled to the open transmitter.

The first circuit serves as impedance converter (see Sect. 5.7) between the energy source (anode voltage source) and the consumer (emitted radiation energy by the oscillating charge in the straight wire).

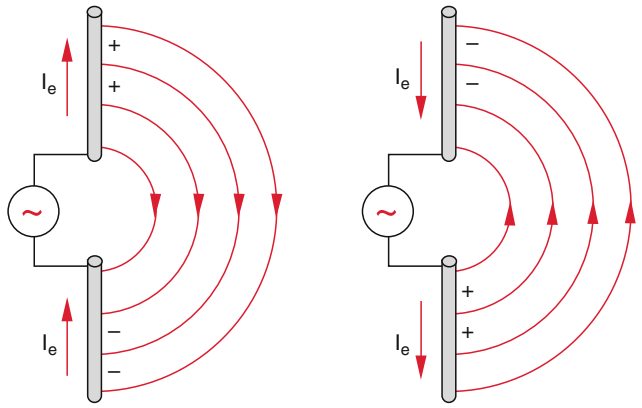


Fig. 6.16 Schematic representation of the generation of high frequency current in a rod antenna. The two drawings show the electron current and a cut through the electric field lines at two phases of the ac generator shifted by 180°. The field lines have rotational symmetry about the antenna

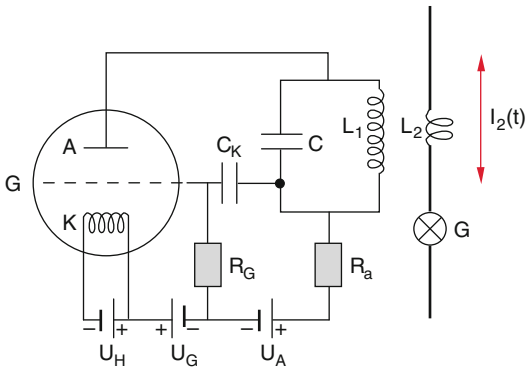


Fig. 6.17 Inductive coupling of an open oscillating circuit with a closed circuit oscillating with constant amplitude induced by capacitive coupling to the grid G of a triode. The inductance L_2 is not necessarily a coil but can be the inductivity of the straight wire

As has been shown in Sect. 5.7 the energy transfer is optimum, if the real resistances of source and consumer are equal and the reactances are opposite equal. If the capacitance of the first circuit with the inductances $L_1 + L_{12}$ is chosen such that the circuit is in resonance with the wanted frequency, its resistance Z becomes real. By the proper choice of L and C the amount $|Z|$ can be most suitably matched with the impedance of the energy generator (electron tube plus resistance R_a). The inductive coupling between the closed oscillation circuit and the antenna acts like a transformer that transforms to the small resistance of the antenna (large current).

The current $I_2(t)$ through the antenna can be visualized by a small light bulb. Any change of the coupling degree k between L_1 and L_{12} (for instance by changing the distance between the two coils or the angle between them) is

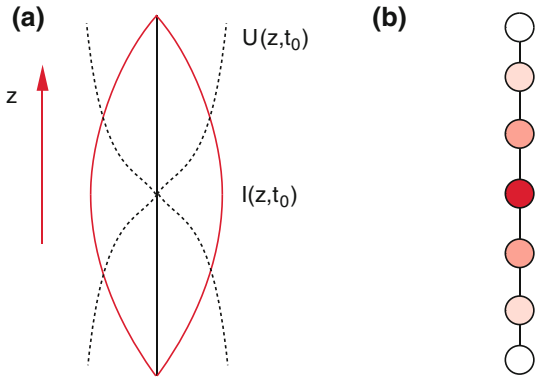


Fig. 6.18 **a**) Current $I(z, t_0)$ and voltage distribution $U(z, t_0)$ along a straight wire at $t = t_0$. **b**) Detection of the current distribution $\langle I^2(z)\rangle$ with small light bulbs

indicated by the changing brightness of the light bulb. Since the brightness is proportional to T^4 (T = temperature of the filament in the light bulb which depends on the electric power $R \cdot \langle I\rangle^2$ consumed in the antenna), already a small change of the current I causes a readily visible change of the brightness.

We assume the antenna with length l to be orientated in the z-direction. If the current through the antenna is described by

$$I(z,t) = I_0(z) \cdot \sin \omega t,$$

The boundary condition

$$I(z = \pm 1/2l) = 0$$

demands that the current amplitude is zero at both ends of the antenna (Fig. 6.18).

The resonant ac-current $I(\omega, t)$ has a spatial amplitude distribution $I_0(z)$, which forms a standing wave with possible wavelengths $\lambda_n = 2l/n$, where l is the length of the antenna and n is integer [3].

The lowest resonance frequency of the antenna is then

$$\omega_0 = \frac{2\pi v_{\mathrm{Ph}}}{\lambda} = \frac{\pi}{l} \cdot v_{\mathrm{Ph}},$$

with the phase velocity v_{ph}

$$v_{\mathrm{Ph}} = \frac{c}{\sqrt{\varepsilon \cdot \mu}} = \frac{1}{\sqrt{\varepsilon\varepsilon_0\mu\mu_0}}$$

This is the velocity of the electro-magnetic field along the antenna. The velocity of light in vacuum is $c = 1/\sqrt{(\varepsilon_0 \cdot \mu_0)}$.

The current distribution along the antenna can be again demonstrated by several small light bulbs placed along the antenna. Their brightness is proportional to $I_0^2(z)$

The current distribution $I(z)$ is shifted against the voltage distribution by $\lambda/4$, because the voltage maxima appear at both ends of the antenna where the charge separation is maximum.

6.4.2 The Electromagnetic Field of the Oscillating Dipole

We regard a conductive straight wire with the charge density ρ. When an ac-current is induced, the freely moving electrons oscillate against the fixed ion cores of the metallic material. The current density $\boldsymbol{j} = \rho \cdot \boldsymbol{v}$ of the electrons depends on the charge density ρ and the velocity $\boldsymbol{v}(t)$ of the oscillating electrons.

According to Eq. (3.14) the vector potential $\boldsymbol{A}(\boldsymbol{r}_1)$ of a stationary current distribution $\boldsymbol{j}(\boldsymbol{r}_2)$ is

$$\boldsymbol{A}(\boldsymbol{r}_1) = \frac{\mu_0}{4\pi} \int_{V_2} \frac{\boldsymbol{j}(\boldsymbol{r}_2)\,\mathrm{d}V_2}{r_{12}}, \tag{6.21}$$

where $r_{12} = |\boldsymbol{r}_1 - \boldsymbol{r}_2|$ is the distance between the charge $\mathrm{d}q = \varrho \cdot \mathrm{d}V_2$ and the observation point $P(\boldsymbol{r}_1)$ (Fig. 6.19).

In order to determine the vector potential $\boldsymbol{A}(\boldsymbol{r}_1, t)$ of a time-dependent current density $\boldsymbol{j}(\boldsymbol{r}_2, t)$ one has to take into account that the propagation of the electromagnetic field, that is generated by the oscillating charge at the position $\boldsymbol{r}_2$ takes the time $\Delta t = (r_1 - r_2)/c$ to reach the observer in $P(\boldsymbol{r}_1)$. Every change of the field in the volume element $\mathrm{d}V_2$ caused by a changing current need the time Δt to arrive in $P(\boldsymbol{r}_1)$ (**retardation**).

Therefore one must consider in (6.21), that the vector potential $\boldsymbol{A}(\boldsymbol{r}_1, t)$ which is measured in $P(\boldsymbol{r}_1)$ at the time t has been generated by currents in the volume $\mathrm{d}V_2$ at the earlier time $(t - r_{12}/c)$. The Eq. (6.21) that was derived for stationary currents has to be modified into

$$\boldsymbol{A}(\boldsymbol{r}_1, t) = \frac{\mu_0}{4\pi} \int \frac{\boldsymbol{j}(\boldsymbol{r}_2, t - r_{12}/c) \cdot \mathrm{d}V_2}{r_{12}}. \tag{6.22}$$

For large distances of the observation point P_1 from the antenna with length l ($r_{12} \gg l$) (6.22) can be readily solved if the following approximations are made:

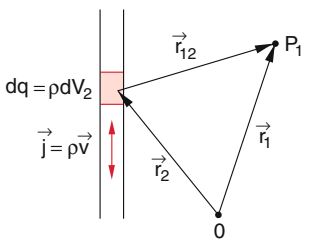


Fig. 6.19 Determination of the time dependent vector potential $\boldsymbol{A}$ in the observation point P_1 generated by the oscillating charge distribution $j = \rho \cdot \boldsymbol{v}(t)$ in the wire

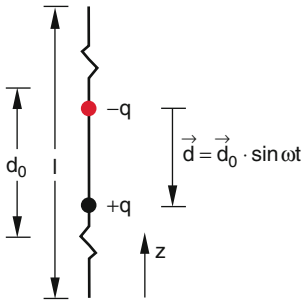


Fig. 6.20 Hertzian dipole

- For a fixed observation point P_1 is the distance to all points of the antenna nearly equal. This means that $1/r_{12}$ can be extracted from the integral.
- The velocity v of the oscillating charge is small compared to the velocity c of light.
- The travel time $\tau = l/c$ of the electromagnetic wave along the length l of the antenna is small compared with the oscillation period $T = 2\pi/\omega$ of the oscillating charge $\mathrm{d}q = \rho \cdot \mathrm{d}V_2$. This implies that the difference $\Delta(r_{12}/c)$ of the travel time $t = r_{12}/c$ from different points of the antenna is small compared with the oscillation period T, i.e. the waves, which are emitted at time t_2 from different points r_2 of the antenna arrive in P_1 nearly at the same time $t_2 = t_1 + r/c$; which means that they all have nearly the same phase.

With these approximations (6.22) converts to

$$\boldsymbol{A}(\boldsymbol{r}_1, t) = \frac{\mu_0}{4\pi r} \int \boldsymbol{v} \cdot \varrho(\boldsymbol{r}_2, t - r/c)\, \mathrm{d}V_2. \tag{6.23}$$

Since the rf-current through the antenna is caused by the flux of electrons with the charge density ρ, we can interpret the integrand in (6.23) as a negative charge $\mathrm{d}q = \rho \cdot \mathrm{d}V_2$ which moves with the time-dependent velocity $v(t)$ against the positive ions in the conductive material of the antenna (Fig. 6.20). The distance d between the centers of positive and negative charge distributions changes as $d = d_0 \cdot \sin\omega t$ when the current through the antenna is $I = I_0 \cdot \cos\omega t$. The time dependent dipole moment $p(t)$ of this Hertzian Dipole is

$$\boldsymbol{p}(t) = q \cdot d_0 \cdot \sin\omega t \cdot \hat{\boldsymbol{e}}_z = q \cdot \boldsymbol{d}. \tag{6.24}$$

Note The amplitude d_0 is much smaller than the length l of the antenna, because the electrons move with the velocity $v \ll c$ and cover during a quarter period $T/4$ of the oscillation only the distance $d_0 = \frac{1}{4} v \cdot T$. However, all N electrons in the antenna participate in the oscillation, i.e. $q = N \cdot e$.

Example

For copper the mobility of the electrons is $u = 4.3 \times 10^{-3}$ (m/s)/(V/m). At an electric field strength $E = 10^3$ V/m → the drift velocity is $v_d = u \cdot E = 4.3$ m/s. At the oscillation frequency $v = 10$ MHz → $T = 10^{-7}$ s $d_0 = ¼ \cdot 4.3 \times 10^{-7}$ m $\approx 10^{-7}$ m, while the length l of the antenna is some meters.

With $v_{ph} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t$ we get from (6.24)

$$\frac{\mathrm{d}\boldsymbol{p}}{\mathrm{d}t} = q \cdot \boldsymbol{v},$$

The vector potential of the Hertzian dipole is then

$$\boldsymbol{A}(\boldsymbol{r}_1, t) = \frac{\mu_0}{4\pi r} \frac{\mathrm{d}}{\mathrm{d}t} \boldsymbol{p}(t - r/c). \tag{6.25}$$

With $\omega \cdot (t - r/c) = \omega t - (2\pi/\lambda) \cdot r = \omega t - kr$ the vector potential becomes

$$\boldsymbol{A}(\boldsymbol{r}_1, t) = \frac{\mu_0}{4\pi} q \cdot d_0 \cdot \omega \frac{\cos(\omega t - kr)}{r} \hat{\boldsymbol{e}}_z. \tag{6.26}$$

This is the equation of a spherical wave (see Vol. 1, Sect. 11.9.4) which starts from the center of the Hertzian dipole and propagates with the velocity $c = \omega/k$ (velocity of light) into the surrounding space.

The oscillating charge q generates an oscillating vector potential $\boldsymbol{A}(\boldsymbol{r}, t)$ and therefore also a magnetic and electric field which propagate into space as electro-magnetic waves with the speed of light c.

What are the properties of the two fields?

For the calculation of the magnetic field we choose the dipole axis as the z-axis (Fig. 6.21). With $\boldsymbol{A} = \{0, 0, A_z\}$ and $\boldsymbol{B} = \mathbf{rot}\, \boldsymbol{A}$ (see Sect. 3.2) we get the relations

$$B_x = \frac{\partial A_z}{\partial y}\,; \quad B_y = -\frac{\partial A_z}{\partial x}\,; \quad B_z = 0\,, \tag{6.27}$$

This shows that the B-field lies in the x-y-plane.

When performing the spatial derivative with respect to y, we must observe that also $\boldsymbol{r}(x, y, z)$ depends on y. Using the chain rule we obtain with $p = |\mathbf{p}| = p_z$

$$B_x = \frac{\mu_0}{4\pi} \left[\dot{p}\left(t - \frac{r}{c}\right) \frac{\partial}{\partial y}\left(\frac{1}{r}\right) + \frac{1}{r} \frac{\partial}{\partial y} \left(\dot{p}\left(t - \frac{r}{c}\right)\right)\right].$$

Using the abbreviation $u = t - r/c$ and $\dot{p} = \mathrm{d}p/\mathrm{d}u$ we get with $\partial u/\partial r = -1/c$, $r = \sqrt{x^2 + y^2 + z^2} \rightarrow \partial r/\partial y = y/r$

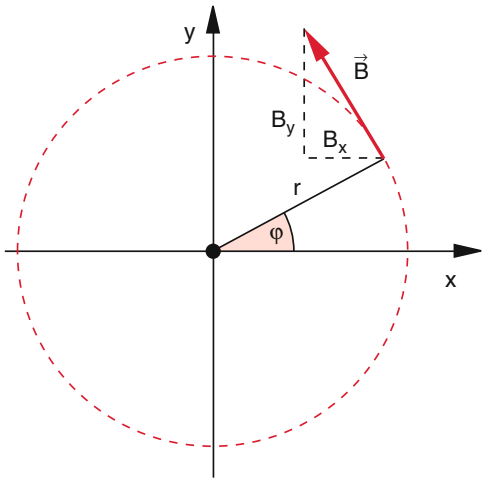


Fig. 6.21 Determination of the magnetic field B from the vector potential of the oscillating dipole with the dipole axis in the z-direction

With

$$\frac{\partial}{\partial y}\left(\frac{1}{r}\right) = -y/r^3$$

We finally obtain

$$B_x = -\frac{1}{4\pi\varepsilon_0 c^2}\left[\dot{p}\frac{y}{r^3} + \ddot{p}\frac{y}{c \cdot r^2}\right]. \tag{6.28a}$$

where we have used the relation $\mu_0 \cdot \varepsilon_0 = 1/c^2$. In a similar way we get

$$B_y = \frac{1}{4\pi\varepsilon_0 c^2}\left[\dot{p}\frac{x}{r^3} + \ddot{p}\frac{x}{c \cdot r^2}\right]. \tag{6.28b}$$

The Cartesian coordinates x and y of the point $P(x, y, z)$ can be transformed into polar coordinates

$$x = r \cdot \sin\vartheta \cdot \cos\varphi\ ; \quad y = r \cdot \sin\vartheta \cdot \sin\varphi,$$

where r is the distance between P and the center of the dipole, which is chosen as the coordinate origin and ϑ is the angle between $\boldsymbol{r}$ and the dipole axis (Fig. 6.22). The Eqs. (6.28a, 6.28b) read in polar coordinates:

$$B_x = -\frac{1}{4\pi\varepsilon_0 c^2}\left[\frac{\dot{p}(u)\sin\vartheta\sin\varphi}{r^2} + \frac{\ddot{p}(u)\sin\vartheta\sin\varphi}{r \cdot c}\right], \tag{6.29a}$$

$$B_y = \frac{1}{4\pi\varepsilon_0 c^2}\left[\frac{\dot{p}(u)\sin\vartheta\cos\varphi}{r^2} + \frac{\ddot{p}(u)\sin\vartheta\cos\varphi}{r \cdot c}\right], \tag{6.29b}$$

We can combine both equations for the components in the vector equation

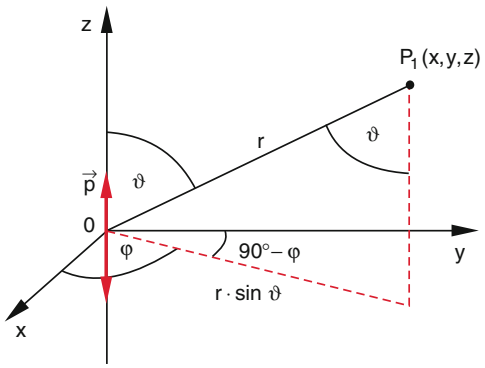


Fig. 6.22 Illustration of the derivation of Eq. 6.29a, 6.29b

$$\boldsymbol{B}(\boldsymbol{r}, t) = \frac{1}{4\pi\varepsilon_0 c^2 r^3}\left[(\dot{\boldsymbol{p}} \times \boldsymbol{r}) + \frac{r}{c}(\ddot{\boldsymbol{p}} \times \boldsymbol{r})\right] \tag{6.30}$$

Note, that because of the retardation the magnetic field $B(\boldsymbol{r}, t)$ is generated by the dipole p at the earlier time $(t - r/c)$. Therefore in (6.30) the quantities $\dot{p}$ and $\ddot{p}$ must be calculated for the time $(t - r/c)$.

Since $\boldsymbol{p} \parallel \dot{\boldsymbol{p}} \parallel \ddot{\boldsymbol{p}}$ it follows that $\boldsymbol{B} \perp \boldsymbol{p}$ and $\boldsymbol{B} \perp \boldsymbol{r}$.

At large distances from the dipole ($r \gg d_0$) the magnetic field $\boldsymbol{B}$ is perpendicular to the dipole axis and perpendicular to the propagation direction $\boldsymbol{r}$ of the electromagnetic wave emitted by the dipole.

The magnetic field (6.30) has two parts, which decrease with different powers of the distance r. At large distances the second term ($\sim 1/r$) where $\ddot{\boldsymbol{p}}$ is dominant because it decreases only with $1/r$. The first term with $\dot{\boldsymbol{p}}$ decreases as $1/r^2$. It is dominant at small distances.

The comparison with the Biot-Savart law (3.16)

$$\mathrm{d}\boldsymbol{B} = \frac{1}{4\pi\varepsilon_0 c^2}\frac{\boldsymbol{j} \times \boldsymbol{r}}{r^3} \cdot \mathrm{d}V$$

shows that because of $\int \boldsymbol{j} \cdot \mathrm{d}V = \dot{\boldsymbol{p}}$ the first term describes the magnetic field, that is directly generated by the oscillating current density $j(t)$.

The second term in (6.30) is indeed indirectly also caused by the oscillating dipole, but the fact that it decreases more slowly with increasing distance than the first term, indicates that an additional source must be present. We will clear this point as follows:

In Fig. 6.23 we regard the oscillating magnetic field B at the point P in the x-y-plane. The radius vector $\boldsymbol{r}$ of $P(\boldsymbol{r})$ is perpendicular to the dipole axis, i.e. $\vartheta = 90^\circ$. The second term in (6.30) then gives the magnetic field in the space-fixed point P where $\dot{\boldsymbol{p}} \perp \boldsymbol{r}$ and $\ddot{\boldsymbol{p}} \perp \boldsymbol{r}$.

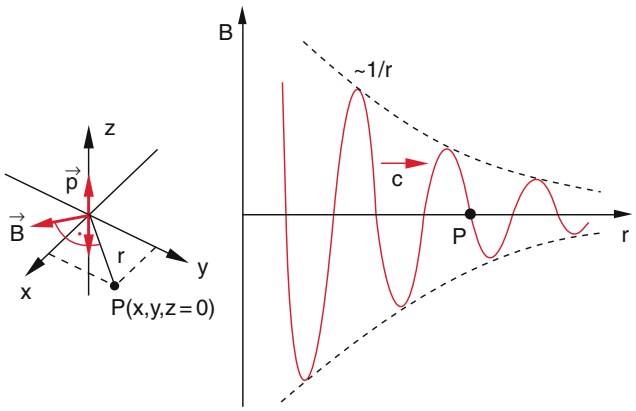


Fig. 6.23 Illustration of the second term in Eq. (6.30) The dashed curve gives the envelope of the spatially with $1/r$ decreasing magnetic field amplitude

$$|\boldsymbol{B}(\boldsymbol{r},t)| = \frac{\ddot{p}}{4\pi\varepsilon_0 c^3 r} = \frac{q d_0 \omega^2}{4\pi\varepsilon_0 c^3 r}\sin(\omega t - kr).$$

While the envelope of the amplitude that decreases with $1/r$ describes the spatial change of B, for the observer in the space-fixed point P the variation of $\dot{B}$ is mainly caused by the high speed c of the electro-magnetic wave passing through P. This means that $\mathrm{d}B/\mathrm{d}t$ is for the observer very large. This large derivative $\mathrm{d}B/\mathrm{d}t$ generates, according to Faraday's law of induction, an alternating electric field $E(t)$ in $P(\boldsymbol{r})$. This in turn creates according to (4.25b) a displacement current which generates an alternating magnetic field

This magnetic field, which is not directly generated by the dipole but by the alternating electric part of the electromagnetic wave is described by the second term in (6.30).

Note, that both parts for the generation of the magnetic field are already included in the Maxwell Eq. (4.25b)

$$\mathbf{rot}\,\boldsymbol{B} = \mu_0 \boldsymbol{j} + \frac{1}{c^2}\frac{\partial \boldsymbol{E}}{\partial t}$$

The first term in (4.25b) corresponds to the first term in (6.30), the second term in (4.25b) to the second term in (6.30).

The oscillating fields $\boldsymbol{E}(\boldsymbol{r}, t)$ and $\boldsymbol{B}(\boldsymbol{r}, t)$ which are produced by the oscillating dipole, propagate with the velocity of light c into the surrounding space. At every point of the space electric and magnetic fields generate each other because of their time variation. These "secondary fields" superimpose the fields that are directly generated by the oscillating dipole. With increasing distance from the dipole the relative contribution of the secondary fields increases, because their amplitude decreases only with $1/r$, whereas the primary wave amplitude declines with $1/r^2$.

The electric field can be deduced from the electric potential ϕ_{el} which is related to the vector potential A by the Lorenz's gauge condition

$$\mathrm{div}\,\boldsymbol{A} = -\frac{1}{c^2}\frac{\partial \phi_{\mathrm{el}}}{\partial t} \tag{6.31}$$

With $\boldsymbol{A} = \{0, 0, A_z\}$ is $\mathrm{div}\,\boldsymbol{A} = \partial A_z/\partial z$ and we can, quite analogue to the calculation of B_x in (6.28a, 6.28b), directly get the derivative of $\boldsymbol{A}$. From (6.25) we then get

$$\nabla \cdot \boldsymbol{A} = -\frac{1}{4\pi\varepsilon_0 c^2}\frac{\boldsymbol{r}\cdot\left[\dot{\boldsymbol{p}} + \left(\frac{r}{c}\right)\ddot{\boldsymbol{p}}\right]_{(t-r/c)}}{r^3}. \tag{6.32}$$

With (6.31) we obtain the electric potential by time-integration

$$\phi_{\mathrm{el}}(\boldsymbol{r},t) = \frac{1}{4\pi\varepsilon_0}\frac{\boldsymbol{r}\cdot\left[\boldsymbol{p} + \left(\frac{r}{c}\right)\dot{\boldsymbol{p}}\right]_{(t-r/c)}}{r^3}, \tag{6.33}$$

which finally gives with (4.28) the electric field

$$\boldsymbol{E} = -\nabla\phi_{\mathrm{el}} - \frac{\partial \boldsymbol{A}}{\partial t}$$

We can again compose $\boldsymbol{E}$ as the sum of two terms:

$$\boldsymbol{E}(\boldsymbol{r},t) = \boldsymbol{E}_1(\boldsymbol{r},t) + \boldsymbol{E}_2(\boldsymbol{r},t). \tag{6.34a}$$

The first term can be calculated from (6.33) with $\boldsymbol{E} = -\mathbf{grad}\ \phi_{\mathrm{el}}$

$$\boldsymbol{E}_1(\boldsymbol{r},t) = \frac{1}{4\pi\varepsilon_0 r^3}\left[-\boldsymbol{p}^* + 3(\boldsymbol{p}^*\cdot\hat{\boldsymbol{r}})\cdot\hat{\boldsymbol{r}}\right] \tag{6.34b}$$

With the abbreviation

$$\boldsymbol{p}^* = \boldsymbol{p}(t - r/c) + \frac{r}{c}\dot{\boldsymbol{p}}(t - r/c) \tag{6.34c}$$

where $\hat{\boldsymbol{r}} = \boldsymbol{r}/r$. Equation (6.34b) describes the electric field of a time dependent electric dipole p^*, if the retardation is taken into account.

The electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ is generated by the electric moment p at the earlier time $(t - r/c)$ and its time derivative $\dot{\boldsymbol{p}}(t - r/c)$, which causes the current through the dipole.

The second term in (6.34a)

$$\begin{aligned}\boldsymbol{E}_2(\boldsymbol{r},t) &= \frac{1}{4\pi\varepsilon_0 c^2 r^3}\left[-\ddot{\boldsymbol{p}}(t - r/c)\times\boldsymbol{r}\right]\times\boldsymbol{r} \\ &= \frac{1}{4\pi\varepsilon_0 c^2 r}\left[\ddot{\boldsymbol{p}}(t - r/c) - (\hat{\boldsymbol{r}}\cdot\ddot{\boldsymbol{p}}(t - r/c))\hat{\boldsymbol{r}}\right]\end{aligned} \tag{6.34d}$$

is that part of the electric field which is generated by the changing magnetic field. It is proportional to the second derivative $\ddot{p}$ of the dipole moment p and $\boldsymbol{E}_2$ is perpendicular to r and to B as can be seen by the comparison with (6.30). While the first term $\boldsymbol{E}_1$ decreases strongly ($\propto 1/r^3$) with increasing distance r the second term $\boldsymbol{E}_2$ declines only with

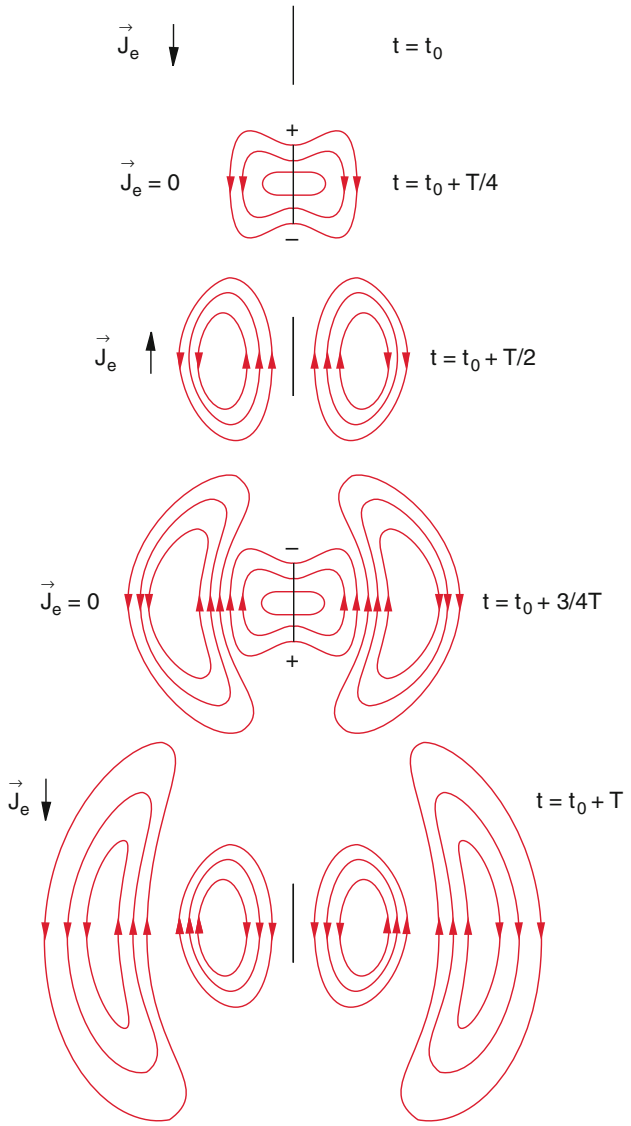


Fig. 6.24 Electric field lines of the Hertzian dipole at times $t = t_0 + n \cdot T/4$. The field lines have cylindrical symmetry about the dipole axis. J_e is the electrical current density in the antenna

$1/r$. In the **near-field range** is $\boldsymbol{E}_1$ dominant, whereas in the **far field range** $\boldsymbol{E}_2$ predominates.

Summarizing we can say:

> The temporally and spatially oscillating electric and magnetic fields represent electro-magnetic waves, which propagate through space with the velocity $v = c$. In each point $P(\boldsymbol{r})$ covered by the wave the changing electric field generates a magnetic field and vice versa.

In a point $P(\boldsymbol{r})$ where r forms the angle ϑ against the dipole axis (Fig. 6.22) Eq. (6.34d) can be written as

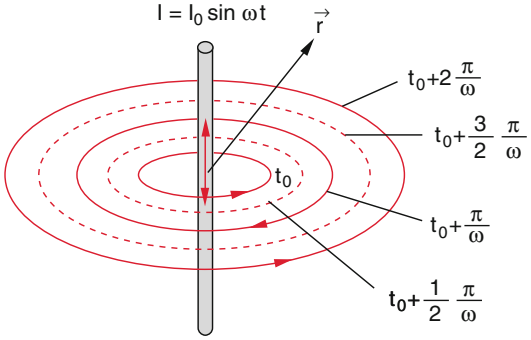


Fig. 6.25 Magnetic field lines of the Hertzian dipole in the equator plane

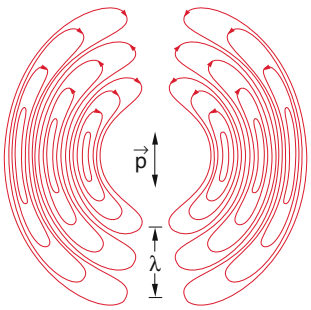


Fig. 6.26 Spatial distribution of the electric field lines. The wavelength λ of the radiated electro-magnetic wave corresponds to twice the spatial distance between the nodes of the electric field

$$|\boldsymbol{E}_2(\boldsymbol{r}, \vartheta, t)| = \frac{\ddot{p}(t - r/c) \sin \upsilon}{4\pi\varepsilon_0 c^2 r}. \tag{6.34e}$$

Note, that the second time derivative of p means $\ddot{p} = \mathrm{d}^2p/\mathrm{d}u^2$ with $u = (t - r/c)$, i.e. the time derivative describes the change of the dipole moment p at the time $(t - r/c)$.

In Fig. 6.24 snapshots of the electric field in the near field range around the dipole are shown at times $t = t_0 + n \cdot T/4$, i.e. every quarter oscillation period.

The magnetic field lines are circles around the dipole axis (Fig. 6.25).

For a given time t_0 and for large distances r the magnetic field amplitude shows the spatial modulation $B(r) = (B_0/r) \cos kr$ with nodes at distances $\Delta r = \pi/k = \pi \cdot c/\omega$.

Analogous statements are valid for the electric field. The electric field in the polar plane of the dipole axis has a kidney-shaped pattern of the field lines (Fig. 6.26). The electric field lines run perpendicularly through the equator plane.

6.5 The Emitted Radiation of the Oscillating Dipole

We have discussed in the previous section that the Hertzian dipole emits radiation in the form of electro-magnetic waves, which propagate into space with the velocity c of light. We will now investigate the emitted power and its frequency spectrum.

6.5.1 The Emitted Power

The comparison of (6.34d) for the electric field and (6.30) for the magnetic field $\boldsymbol{B}$ reveals that at large distances r from the dipole (far field) the amount of $\boldsymbol{B}$ is smaller by the factor $1/c = 3.3 \times 10^{-9}$ s/m than the amount of $\boldsymbol{E}$. Inserting this relation into the Eq. (4.20a) for the energy density of the electro-magnetic field we obtain

$$w_m = \frac{1}{2}\varepsilon_0(E^2 + c^2B^2) = \varepsilon_0 E^2. \tag{6.35}$$

This gives the energy flux (energy which is transported per unit time through the unit area)

$$S = \varepsilon_0 \cdot c \cdot E^2. \tag{6.36a}$$

Inserting for the amount of $\boldsymbol{E}$ the expression (6.34e) and for the dipole moment $p = q \cdot d_0 \cdot \sin(\omega t - r/c) \rightarrow \ddot{p} = -qd_0 \cdot \omega^2 \cdot \sin(\omega t - r/c)$ we obtain in the far field ($r \gg d_0$) the energy that passes per sec through the surface of a sphere with radius r in the direction ϑ against the dipole axis

$$S = \frac{q^2 d_0^2 \omega^4 \sin^2\vartheta}{16\pi^2\varepsilon_0 c^3 r^2}\sin^2\omega((t - r/c)). \tag{6.36b}$$

The dipole emission is maximum in the direction perpendicular to the dipole axis ($\vartheta = 90^\circ$), whereas in the direction of the dipole axis ($\vartheta = 0$) no energy is emitted (Fig. 6.27).

Since the energy flux S is proportional to $1/r^2$ the total flux through the surface of a sphere with radius r is independent of r.

With increasing distance the terms proportional to $1/r$ of the electric field (6.34d) and of the magnetic field (6.30), win more and more importance for the energy transport, whereas the other parts ($\propto 1/r^3$ for $\boldsymbol{E}$ and $\propto 1/r^2$ for $\boldsymbol{B}$) approach faster zero and can be neglected for large distances from the dipole.

The power, transported through the area $dA = r^2 \cdot \sin\vartheta \cdot d\vartheta \cdot d\varphi$ (see Vol. 1, Sect. 13.2.3) is $P = S \cdot dA$. Integration over ϑ and φ gives the total power, emitted by the oscillating dipole into the whole space

$$P_{em} = \oint \boldsymbol{S} \cdot d\boldsymbol{A} = \frac{q^2 d_0^2 \omega^4}{6\pi\varepsilon_0 c^3}\sin^2(\omega(t - r/c)) \tag{6.37}$$

With $\overline{sin^2\omega(t - r/c)} = 1/2$ we obtain for the average power, emitted by the dipole with the oscillation amplitude $p_0 = q \cdot d_0$ oscillating at the frequency $\nu = \omega/2\pi$

$$\langle P_{em}\rangle = \frac{q^2\omega^4 d_0^2}{12\pi\varepsilon_0 c^3}. \tag{6.38}$$

Note that P is proportional to ω^4 [4].

6.5.2 Radiation Damping

The total mechanical energy (potential + kinetic energy) of a harmonic oscillator with mass m, oscillation frequency ω and amplitude d_0 is (see Vol. 1, Sect. 11.6)

$$\bar{W} = \bar{E}_{kin} + \bar{E}_{pot} = \frac{1}{2}m\omega^2 d_0^2. \tag{6.39}$$

This is also true for the Hertzian dipole, where charges q with mass m oscillate with the velocity $v = \omega \cdot d_0 \cdot \cos\omega t$.

If the energy loss of the oscillating dipole is not supplied by external sources the oscillation amplitude will decrease in the course of time due to the emitted radiation energy (6.38). The relative energy loss is given by the ratio of (6.38) and (6.39).

$$\frac{\overline{dW/dt}}{\bar{W}} = -\frac{q^2\omega^2}{6\pi\varepsilon_0 mc^3} = -\gamma. \tag{6.40}$$

From $\overline{dW/dt} = -\gamma\bar{W}$ we obtain by integration

$$\bar{W}(t) = \bar{W}_0 \cdot e^{-\gamma t}. \tag{6.41}$$

After the time $\tau = 1/\gamma$ the energy of the oscillating dipole has dropped to $1/e$ of its initial value $\bar{W}_0 = \bar{W}(t = 0)$ (Fig. 6.28a)

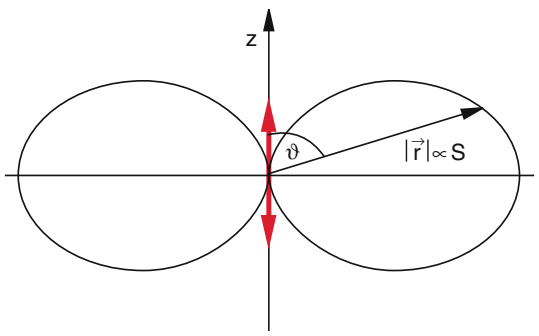


Fig. 6.27 Angular dependence of the radiated power of an oscillating dipole. The length $|r(\vartheta)|$ is proportional to the energy flux density S

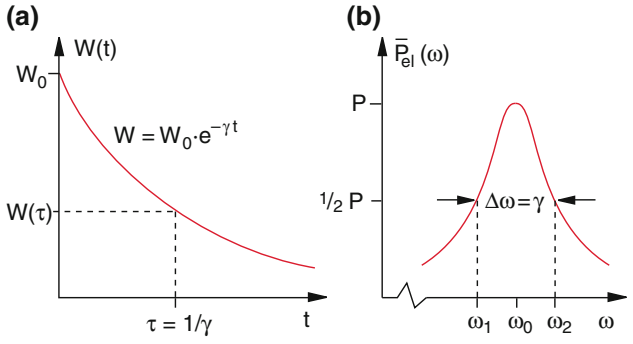


Fig. 6.28 **a**) Exponential decay of the energy of the damped oscillator **b**) frequency spectrum of the radiated power

Example

We can describe an excited atom with an electron of mass m_e by the model of a damped oscillator that releases its excitation energy by emitting light. When we insert into (6.40) the numerical values $m = m_e = 9 \times 10^{-31}$ kg, $q = -e = -1.6 \times 10^{-19}$ C, $\omega = (2\pi c)/\lambda \approx 3.8 \times 10^{15}\,s^{-1}$ for $\lambda = 500$ nm we obtain $\gamma = 9 \times 10^7 s^{-1}$. This gives a damping time $\tau = 1/\gamma = 1.1 \times 10^{-8}$ s.

The mean energy of the excited atom is $\bar{W} \approx 4 \times 10^{-19}$ J. This gives the oscillation amplitude $d_0 = 8 \times 10^{-11}$ m of the excited electron. The mean radiation power of the atom is then

$$\overline{(dW/dt)} = -\gamma \bar{W} = -9 \times 10^7 \cdot 4 \times 10^{-19}\,W \approx 3.6 \times 10^{-12}\,W.$$

In a gas discharge lamp which delivers 1 W visible radiation therefore about 3×10^{11} atoms are excited per s.

6.5.3 Frequency Spectrum of the Emitted Radiation

The oscillation amplitude of a damped oscillator is

$$z = d = d_0 e^{-\beta t} e^{i\omega t},$$

If the oscillator is driven by the electric field strength $E = E_0 \cdot e^{i\omega t}$ it performs forced oscillations which are described by the equation of motion (see Vol. 1, Sect. 11.5)

$$\ddot{z} + 2\beta\dot{z} + \omega_0^2 z = \frac{q}{m} E_0 e^{i\omega t} \tag{6.42}$$

The energy $W \propto d^2$ decays as $W(t) = W_0 e^{-\gamma t}$ if $\beta = \gamma/2$. Inserting the ansatz

$$z = z_0 \cdot e^{i\omega t}$$

Into (6.42) we get the complex oscillation amplitude

$$z_0 = \frac{(q/m)E_0}{(\omega_0^2 - \omega^2) + i\gamma\omega}, \tag{6.43}$$

with the square of the absolute value

$$|z_0|^2 = \frac{(q^2/m^2)E_0^2}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2} \tag{6.44}$$

With $|z_0| = d_0$ we get from (6.38) the frequency spectrum of the mean radiation power (Fig. 6.38b)

$$\bar{P} = \frac{d\bar{W}}{dt} = \frac{q^4\omega^4 E_0^2}{12\pi\varepsilon_0 m^2 c^3} \frac{1}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}. \tag{6.45}$$

For $(\omega_0^2 - \omega^2)^2 = \omega^2\gamma^2$ the power drops to ½ of its maximum value at $\omega = \omega_0$. This gives the two solutions

$$\omega_{1,2} = \sqrt{\omega_0^2 + \gamma^2/4} \pm \gamma/2$$

For those frequencies where the power has decreased to ½ of its maximum value. The frequency interval $\Delta\omega = \omega_1 - \omega_2 = \gamma$ is therefore called the *full width at half maximum* (FWHM) of the emitted radiation.

Example

When light of the correct frequency passes through an ensemble of atoms they can absorb the light and are excited into energetically higher states. The excitation energy is subsequently emitted as resonance fluorescence, where the time dependence follows (6.41). Changing the frequency of the exciting light the total fluorescence power follows Eq. (6.45). With $\gamma = 10^8\,s^{-1}$ and $\omega = 3.8 \times 10^{15}\,s^{-1}$ the spectral half-width becomes $\Delta\omega = 10^8\,s^{-1} \rightarrow \Delta\nu = 16$ MHz. The relative linewidth is then with $\Delta\omega/\omega = \gamma/\omega = 2.6 \times 10^{-8}$ very small. Excited atoms emit their radiation only within very small frequency intervals.

6.5.4 The Radiation of an Accelerated Charge

We have seen in Sect. 6.4.2 that the amplitude E_0 of the electromagnetic wave emitted by the oscillating dipole, is in the far field ($r \gg d_0$) proportional to the second derivative $\ddot{p}$ of the dipole moment $p = q \cdot d$, i.e. to the acceleration $a = \ddot{d}$ of the oscillating charge q. The emitted radiation power is then proportional to the square a^2 of the acceleration.

This statement is not restricted to harmonic oscillations but is valid quite general for arbitrary acceleration of charges [4].

The following discussion illustrates the form of the electromagnetic waves emitted by accelerated charges [5].

In Sect. 3.4.1 we have discussed the electric field of a charge moving with the velocity v. When the charge is accelerated, the velocity v changes either its amount or its direction or both. This changes the spatial distribution of the electric field. This is illustrated again in Fig. 6.29a–d.

Figure 6.29a shows the electric field lines of a charge q at rest. If q is accelerated at the time $t = t_0$ nearly abruptly to a large velocity $v \approx c$ the electric field pattern changes to that of a moving charge (Fig. 6.29b). This change cannot be present immediately in the whole space but propagates with the velocity c of light. The modified field generated by the charge at the time $t_1 = t_0 + \Delta t$ in the point B cannot be observed by an observer at the time t_2, if his distance from

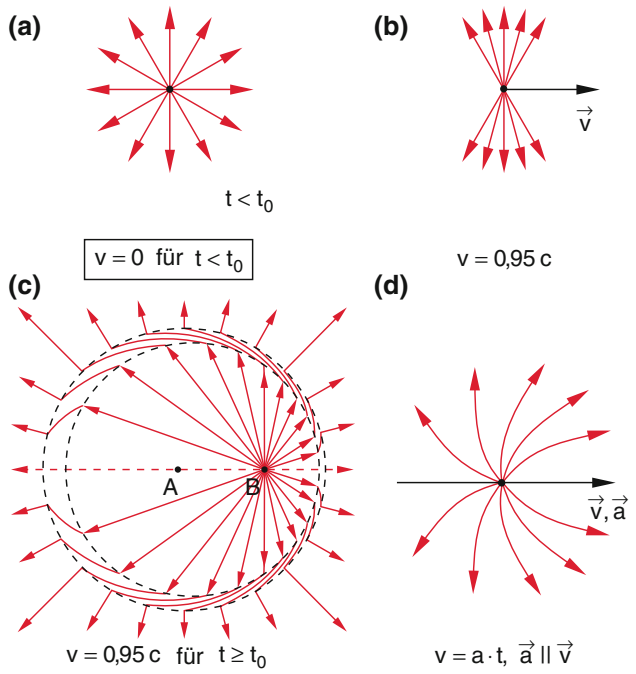


Fig. 6.29 **a**) Electric field lines of a charge at rest, **b**) stationary field lines of a charge moving with constant velocity v, **c**) field lines of a charge q at time $t = t_0 + R/c$ if the charge at rest had been suddenly accelerated at time t_0 to the velocity v, **d**) field lines of a continuously accelerated charge with $\boldsymbol{a} \parallel \boldsymbol{v}$

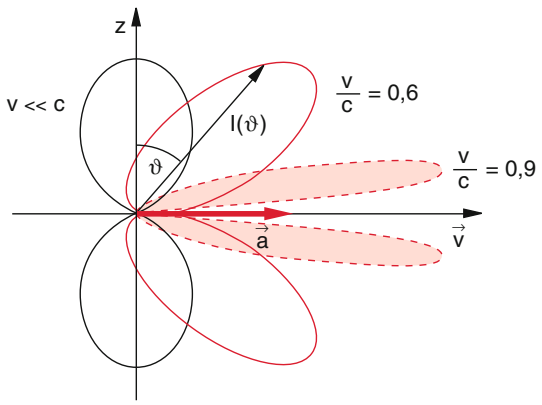


Fig. 6.30 Cut through the angular distribution $I(\vartheta)$ of the radiation power, which has rotational symmetry about the direction of the acceleration $\boldsymbol{a} \parallel \boldsymbol{v}$ of a moving charge for different velocities

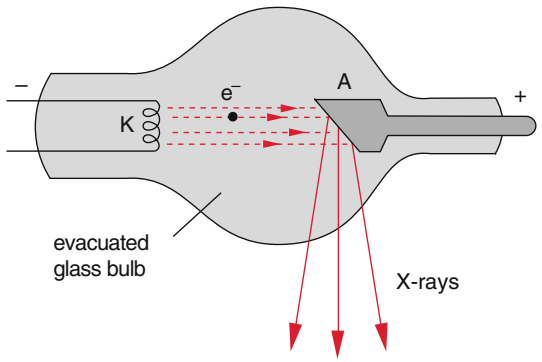


Fig. 6.31 X-ray tube

the source is larger than $c \cdot (t_2 - t_1)$. He then still observes the field of a charge resting in A.

Since the field lines of a point charge at rest are equally distributed over all directions (Coulomb field) but are compressed for a moving charge around the angle $\alpha = 90°$ against the direction of $\boldsymbol{v}$, there is a sudden jump on the surface $R = c(t_2 - t_1)$. It is shown schematically in Fig. 6.29c.

For the more realistic case of a uniform acceleration the change of the field line pattern does not occur abruptly but continuously. For a uniform acceleration of a charge q one obtains instead of the sudden jump a curvature of the field lines (Fig. 6.29d) (see for instance the Ealing teaching movie "*charges that start and stop*" [5].

A similar situation occurs for the magnetic field. When the velocity of the charge changes, the current density $j = q \cdot v$ changes correspondingly and therefore also the magnetic field.

The emitted power of a charge that moves with the velocity v parallel to the acceleration $\boldsymbol{a}$, shows an angular distribution, which is tilted towards the direction of the acceleration away from the dipole axis (Fig. 6.30).

The general treatment of the radiation of charges that are accelerated in an arbitrary way can be found in textbooks of theoretical electrodynamics. In the present textbook, we will restrict the discussion to two examples.

6.5.4.1 Bremsstrahlung of X-Rays

In an evacuated tube electrons are emitted by a hot cathode K (Fig. 6.31) and are accelerated to large velocities by a high anode voltage of about 10–100 keV. These energetic electrons hit the cathode, made of copper or tungsten. The electrons are deflected in the Coulomb field of the atomic nuclei (Fig. 6.32). This change of the direction of v over an extremely short distance represents a large acceleration and results in the emission of a continuous radiation with a broad spectrum in the X-ray region (**bremsstrahlung**)

6.5.4.2 Synchrotron Radiation

Electrons that have been accelerated to very high energies (MeV–GeV) and velocities close to the velocity c of light can be forced by a magnetic field on a circular path with radius R, where the Lorentz force $F_L = e \cdot (\boldsymbol{v} \times \boldsymbol{B})$ and the centrifugal force $F_c = m \cdot v^2/R$ just cancel. The acceleration of the electrons, which move with constant velocity v around the ring is $a = v^2/\boldsymbol{R}$. The vector $\boldsymbol{a}$ is always perpendicular to $\boldsymbol{v}$. The emitted radiation is proportional to a^2. For large velocities the spatial distribution of the emitted radiation is strongly concentrated around the velocity v (Fig. 6.33b).

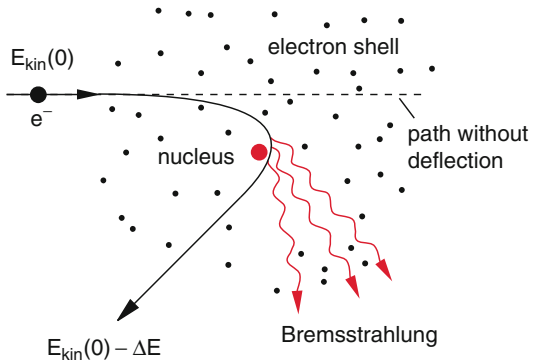


Fig. 6.32 Deceleration of electrons in the Coulomb field of the atomic nuclei in the anode

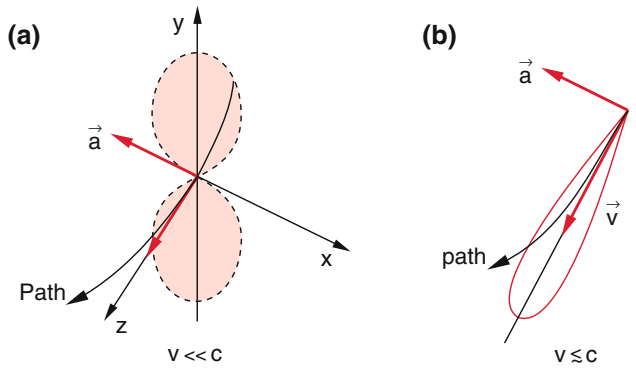


Fig. 6.33 Radiation characteristic of an accelerated charge that moves with constant velocity on a circular path **a**) for $v \ll c$ the distribution has rotational symmetry about the x-axis = direction of the acceleration $\boldsymbol{a}$. **b**) With increasing velocity $\boldsymbol{v}$ the distribution becomes more and more peaked within a narrow angular range around the tangent to the circular path

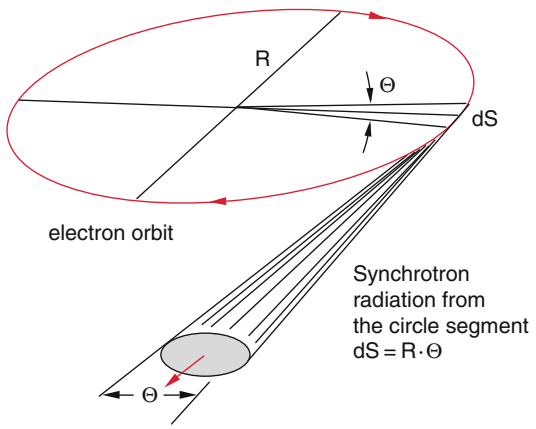


Fig. 6.34 Synchrotron radiation emitted by electrons running with constant velocity amount on a circular path

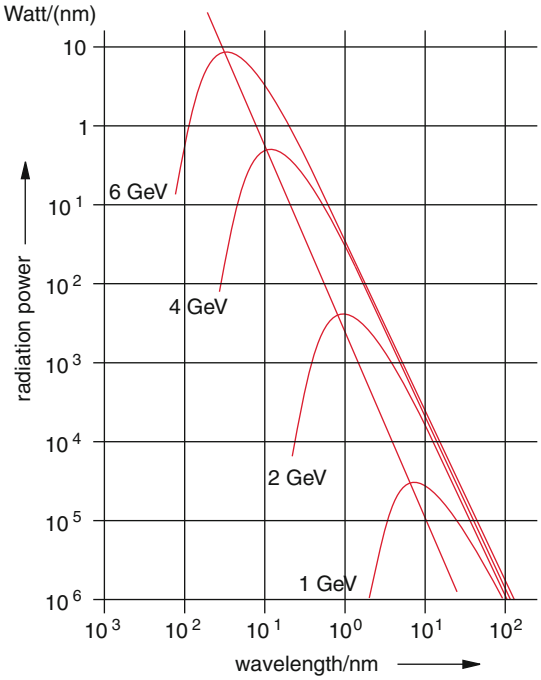


Fig. 6.35 Spectral distribution of the synchrotron radiation within a spectral interval of $\Delta\lambda = 1$ nm in the storage ring DORIS for an electron beam of 0.16 pA (10^6 e/s) for different electron energies [8]

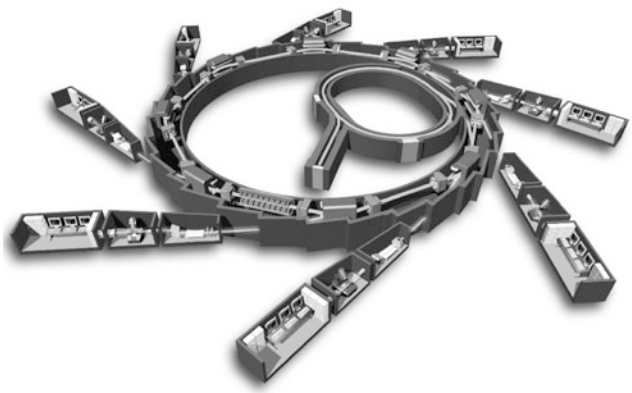


Fig. 6.36 Synchrotron SOLEIL in Gif-sur-Yvette near Paris with 8 beam lines and measuring stations The small ring in the inner part is the pre-accelerator for the electrons

The electrons in a synchrotron have velocities $v \approx 0.99999c$ and their emitted radiation is therefore concentrated within a narrow angular range around the tangent to the electron path (Fig. 6.34).

In Fig. 6.35, the spectral distributions of the synchrotron radiation of the storage ring DORIS in Hamburg are shown [6] for several values of the electron energy. This illustrates that for example at $W = 6$ GeV the maximum of the distribution occurs at the wavelength $\lambda = 0.03$ nm, i.e. in the X-ray region. Generally many beamlines are installed tangentially to the circular path of the electrons. In Fig. 6.36 the 8 beam lines of the synchrotron SOLEIL near Paris are illustrated [7].

Summary

- Electromagnetic oscillations in an oscillation circuit of capacitor and inductance constitute a periodic exchange of electric energy in the charged capacitor and magnetic energy in the inductance coil.
- The resonant frequency of the oscillation in the circuit consisting of capacitor C, Inductance L and Ohmic resistor R is

$$\omega = \sqrt{\frac{1}{LC} - \frac{R^2}{4L^2}}.$$

- The oscillation energy can be transferred from one oscillation circuit to another, coupled to the first one by inductive, capacitive or Ohmic coupling. The degree of coupling is for inductive coupling $k = L_{12}/\sqrt{L_1 \cdot L_2}$.
- For the open oscillation circuit the electric and the magnetic fields are no longer localized but spread out as electromagnetic waves into the surrounding space.
- A model for an open oscillation circuit is the Hertzian dipole, where a negative charge $-q$ oscillates periodically against a positive charge $+q$. This causes an oscillating electric dipole moment $p = q \cdot d_0 \cdot \sin \omega t$.
- The electromagnetic power, emitted by the Hertzian dipole into the whole surrounding space is

$$P_{\text{em}} \propto q^2 d_0^2 \omega^4.$$

- The power emitted into the solid angel $\mathrm{d}\Omega$ under the angle ϑ against the dipole axis is for a dipole at rest proportional to $\sin^2 \vartheta \cdot \mathrm{d}\Omega$.
- Every accelerated charge q emits energy in form of electromagnetic waves. The emitted power is $P_{\text{em}} \propto q^2 \cdot a^2$, where a is the amount of the acceleration.
- For large velocities of the charge q ($v \approx c$) amount and angular distribution of the emitted radiation power change with increasing v. They are more and more concentrated in a narrow angular range $\Delta\vartheta$ around the direction of the velocity v. It is $\Delta\vartheta \propto 1/\gamma$ with $\gamma = (1 - v^2/c^2)^{-1/2}$

Problems

6.1. A parallel oscillation circuit oscillates at a frequency ν = 800 kHz. After 30 oscillation periods the voltage amplitude across the capacitor has dropped to 1/2 of its initial value. How large are L and R?

6.2. To which fraction of the maximum value $P(\omega_0)$ has the power $P(\omega)$ in a series oscillating circuit dropped at the frequencies $\omega_1 = \omega_0 \pm R/L$ and $\omega_2 = \omega_0 \pm 2RL$? What is the ratio $|Z(\omega_0 \pm R/L)|/|Z(\omega_0)|$ in the parallel circuit? Why does the maximum of the active power occur at ω_r but not at the resonance frequency ω_0 of the lossless circuit?

6.3. What are the resonant frequencies ω_1 and ω_2 in a system of two coupled equal oscillating circuits with $\omega_0 = 10^6$ s^{-1}, $L = 10^4$ H and $L_{12} = k \cdot L$ with $k = 0.05$?

6.4. The electron in the classical model of the hydrogen atom has a kinetic energy of 13.6 eV and moves on a circle with the radius $R = 5.3 \times 10^{-11}$ m
What would be in a classical model the radiation power
(a) For one revolution and
(b) per second?
(c) How would the path look like, if this energy loss is taken into account? How much would the radius R change per revolution? How long would it take before the electron arrives at the proton?

6.5. What is the radiation power emitted by a charge q which moves with a velocity $v \ll c$ in a plane perpendicular to a magnetic field B? What is the initial radius R of its circular path and what is the change of the velocity v and the radius R in course of time.

6.6. A proton travels in a linear accelerator a distance of 3 m with a potential difference of 10^6 V. It therefore experiences a constant acceleration
(a) Which radiation power does the proton emit?
(b) Compare this with the power, emitted by a proton moving with the energy of 10^6 eV on a circle with circumference of 3 m.
(c) What is the total energy emitted by the proton in (a)?

6.7. A system of microscopic oscillating dipoles, which are concentrated in a tiny volume, emit isotropically a radiation power of 10^4 W
(a) What are the amplitudes of the electric and the magnetic fields at a distance r = 1 m ($r \gg$ diameter of the source)?
What is the intensity of the electromagnetic wave?

6.8. A nonisotropic emitter radiates electromagnetic waves into the solid angle $d\Omega = 10^{-2}$. At a distance of 10^3 m the electric field has the amplitude of 10 V/m. What is the radiation power of the emitter?

6.9. The earth receives from the sun the radiation power density of 1.4×10^3 W/m^2 (solar constant)
(a) What are the electric and magnetic field strengths at the surface of the earth, if reflection and absorption in the atmosphere are neglected?
(b) What is the total power, the sun radiates into all directions?
(c) What is the electric field strength of the radiation at the surface of the sun $R = 6.96 \times 10^8$ m, if other contributions to electric fields are neglected?

6.10. A light bulb with an electric input power of 100 W converts about 70% of this power into isotropic radiation. What is the electric field strength at a distance of 1 m? Compare this with the field strength of the sun radiation. Which input power must the light bulb have in order to generate the same field strength as the sun radiation?

References

1. L. Manewitch, A. Kovaleva, V.Smirnov, Y. Starosvetski: Two Coupled Oscillators. (Springer, Singapore 2018)
2. https://en.wikipedia.org/wiki/Klystron
3. https://en.wikipedia.org/wiki/Dipole_antenna
4. http://www.tapir.caltech.edu/~teviet/Waves/empulse.html
5. NCSU Physics Demonstrations https://www.academia.edu/25764528/RADIATION_OF_AN_ACCELERATED_CHARGE
6. Deutsches Elektronen-Synchrotron DESY Hamburg, storage ring DORIS
7. Synchrotron SOLEIL L'Orme des Merisiers Saint-AubinBP 48 91192 Gif-sur-Yvette Cedex
8. http://photon-science.desy.de/research/students__teaching/primers/synchrotron_radiation/index_eng.html

Electromagnetic Waves in Vacuum

7

In the previous chapter it has been shown, that oscillating dipoles emit electromagnetic radiation. In this chapter we will treat the description and properties of waves in more detail. The reader is advised to consult the analog description for mechanical waves in Vol. 1, Chap. 11.

7.1 The Wave Equation

We start with Maxwell's equations in vacuum without charges and currents ($\rho = 0, j = 0$)

$$\nabla \times \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}, \tag{7.1a}$$

$$\nabla \times \boldsymbol{B} = \varepsilon_0 \cdot \mu_0 \cdot \frac{\partial \boldsymbol{E}}{\partial t}. \tag{7.1b}$$

Now we apply on both sides of (7.1a) the differentiation operator *curl* and insert *curlB* from (7.1b). We get

$$\begin{aligned} \nabla \times \nabla \times \boldsymbol{E} &= -\nabla \times \frac{\partial \boldsymbol{B}}{\partial t} = -\frac{\partial}{\partial t}(\nabla \times \boldsymbol{B}) \\ &= -\varepsilon_0 \cdot \mu_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}, \end{aligned} \tag{7.2}$$

Here we have used the fact that the differentiation with respect to time can be preponed because ∇ does not depend on time.

Now we use the vector relation for **rot rot** $\boldsymbol{E}$ (see Vol. 1, Sect. 13.1.6)

$$\begin{aligned} \nabla \times \nabla \times \boldsymbol{E} &= \nabla(\nabla \cdot \boldsymbol{E}) - \nabla \cdot (\nabla \boldsymbol{E}) \\ &= \mathbf{grad}(\mathrm{div}\, \boldsymbol{E}) - \mathrm{div}(\mathbf{grad}\, \boldsymbol{E}). \end{aligned}$$

In a space without charges is the charge density $\rho = 0$ and therefore according to (1.10) $\mathrm{div}\, \boldsymbol{E} = \rho/\varepsilon_0 = 0$. Therefore we obtain from (7.2) the equation

$$\Delta \boldsymbol{E} = \varepsilon_0 \mu_0 \frac{\partial^2 \boldsymbol{E}}{\partial t^2}, \tag{7.3}$$

where Δ = div **grad** is the Laplace-operator. A comparison with (11.69) in Vol. 1 shows, that (7.3) describes a wave equation for the propagation of a time-dependent electric field $\boldsymbol{E}(\boldsymbol{r}, t)$ in vacuum which propagates at the speed of light

$$c = \frac{1}{\sqrt{\varepsilon_0 \mu_0}} \tag{7.4}$$

This is a vector equation that represents three component equations. As example, the Eq. (7.3) reads for the component E_x in Cartesian coordinates

$$\frac{\partial^2 E_x}{\partial x^2} + \frac{\partial^2 E_x}{\partial y^2} + \frac{\partial^2 E_x}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 E_x}{\partial t^2}. \tag{7.3a}$$

Corresponding equations are valid for the components E_y and E_z.

An analogue wave equation can be derived for the magnetic field $\boldsymbol{B}(\boldsymbol{r}, t)$ if one takes *curlcurl* of (7.1b) and uses (7.1a), (see Problem 7.1).

Note In the SI-system the speed of light can be expressed by (7.4) with the permittivity of vacuum ε_0 and the permeability of vacuum μ_0. This follow from

(a) the wave Eq. (7.3) derived from the Maxwell equations and the comparison with (11.69) in Vol. 1
(b) the comparison of the Lorentz forces in two different systems of inertia (Sect. 3.4.3).

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_7

7.2 Electro-magnetic Plane Waves

Especially simple solutions of the wave Eq. (7.3) are obtained if E depends only on one coordinate, e.g. the z-coordinate.

$$\frac{\partial \boldsymbol{E}}{\partial x} = \frac{\partial \boldsymbol{E}}{\partial y} \equiv \boldsymbol{0}, \tag{7.5}$$

i.e. the vector E has at a fixed time $t = t_0$ on a plane $z = z_0$ everywhere the same magnitude and the same direction. The wave Eq. (7.3) simplifies to

$$\frac{\partial^2 \boldsymbol{E}}{\partial z^2} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}. \tag{7.6}$$

From div $\boldsymbol{E} = 0$ in a space without charges we find from (7.5)

$$\frac{\partial E_Z}{\partial z} = 0 \Rightarrow E_z = a = \text{spatially constant} \tag{7.6a}$$

We choose the boundary conditions so that the constant becomes zero, $a = 0$. The wave has now only the components E_x and Ey.

$$\boldsymbol{E} = \{E_x, E_y, 0\}.$$

The general solutions of (7.6) for plane waves are

$$\begin{aligned} E_x(z,t) &= f_x(z-ct) + g_x(z+ct), \\ E_y(z,t) &= f_y(z-ct) + g_y(z+ct). \end{aligned} \tag{7.7}$$

Here f and g are arbitrary, but continuously differentiable functions with their arguments $(z - ct)$ or $(z + ct)$ (see Vol. 1, Sect. 11.9). They represent plane waves (Fig. 7.1) because the planes z = constant are areas of constant phase. That means, for every point in the plane $z = z_0$ the argument $(z \pm ct)$ is equal at equal times. These phase areas $z = z_0$ move for $f(z - ct)$ with the speed c in the $+z$-direction, because from the phase condition $(z - ct)$ = constant we get by differentiation

$$\frac{\mathrm{d}z}{\mathrm{d}t} - c = 0 \Rightarrow \frac{\mathrm{d}z}{\mathrm{d}t} = +c.$$

For the function $g(z + ct)$ the waves move in the $-z$-direction. The solutions (7.7) of the wave Eq. (7.6) are plane transverse waves because the electric field vector $\boldsymbol{E} = \{E_x, E_y, 0\}$ is perpendicular to the direction of propagation $\boldsymbol{e}_z$.

Note

(a) The transversality $\boldsymbol{E} \perp \boldsymbol{e}_z$. follows from div $\boldsymbol{E} = 0$ and is therefore only valid in charge-free space! In matter with a charge density $\rho \neq 0$ or if conducting boundaries exist, the wave need not be transverse. Examples are waves within wave guides or in anisotropic materials (see Sect. 7.9). The transversallity is in general not given if the wave travels through a space that is limited on both sides. An example is a linear polarized wave traveling into the z-direction through a space that is restricted in the x-direction.

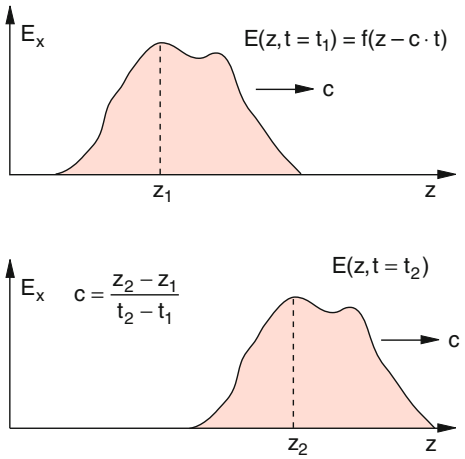


Fig. 7.1 Non-periodical plane wave, propagating into the $+z$-direction

(b) $$\boldsymbol{E}(x,z) = \begin{Bmatrix} E_x \\ 0 \\ -(\mathrm{i}/k)\frac{\partial E_x}{\partial x} \end{Bmatrix}$$

(c) A wave need not be periodic. Think about shock waves (Vol. 1, Sect. 11.13) or electromagnetic pulses, which can be produced by pulsed arcs. They have a broad frequency spectrum with statistically distributed phases of their components. Also these nonperiodic waves are solutions of the wave equation (7.3) and if they follow Eq. (7.7) they are also plane waves.

7.3 Periodic Waves

An important and frequently found special case of electromagnetic waves are periodic plane waves, which can be described by sine- or cosine-functions.

We denote as the wavelength λ the distance between two equal values of the function f in (7.7) at the same time (Fig. 7.2a).

$$f(z + \lambda - ct) = f(z - ct). \tag{7.8}$$

For periodic waves we use the ansatz

$$\boldsymbol{E} = \boldsymbol{E}_0 \cdot f(z - ct) = \boldsymbol{E}_0 \cdot \sin k(z - ct), \tag{7.9a}$$

Then we get for the constant k with the conditions for periodicity (7.8),

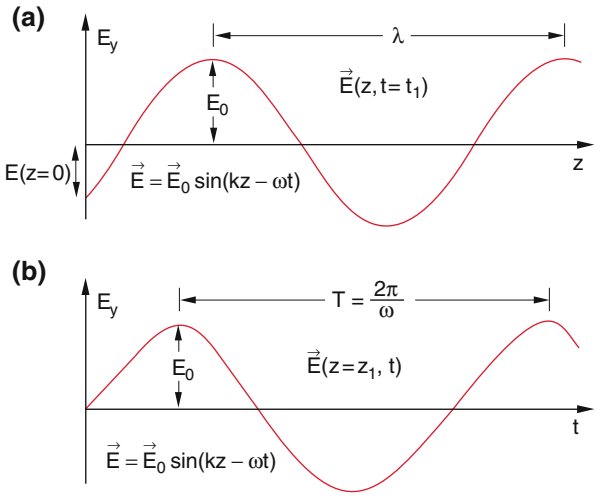


Fig. 7.2 Harmonic electro-magnetic wave with the electric field vector in y-direction propagating into the z-direction **a**) momentary state E(z, $t = t_1$), **b**) time dependence at a given location $z = z_1$

$$k \cdot \lambda = 2\pi \Rightarrow k = \frac{2\pi}{\lambda}. \tag{7.10a}$$

The constant k is named **wave number**. We can rewrite (7.9a), using $c = v \cdot \lambda$, as

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{E}_0 \cdot \sin\left(kz - \frac{2\pi c}{\lambda} t\right) \\ &= \boldsymbol{E}_0 \cdot \sin(kz - \omega t). \end{aligned} \tag{7.9b}$$

Of course we can also apply cosine-functions as periodic solution

$$\boldsymbol{E} = \boldsymbol{E}_0 \cdot \cos(kz - \omega t). \tag{7.9c}$$

The correct choice depends on the initial conditions.

Often a notation with complex numbers is used,

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{A}_0 \, \mathrm{e}^{\mathrm{i}(kz-\omega t)} + \boldsymbol{A}_0^* \, \mathrm{e}^{-\mathrm{i}(kz-\omega t)} \\ &= \boldsymbol{A}_0 \, \mathrm{e}^{\mathrm{i}(kz-\omega t)} + \text{c.c.} \end{aligned} \tag{7.9d}$$

where c.c. stands for the complex conjugate.

If the amplitude A_0 is a real valued vector, (7.9d) becomes

$$\boldsymbol{E} = 2\boldsymbol{A}_0 \cos(kz - \omega t). \tag{7.9e}$$

The comparison with (7.9c) shows, that $\boldsymbol{E}_0 = 2\boldsymbol{A}_0$.

In the shorthand notation of (7.9d) the complex conjugate term is left out. However, always keep in mind, that the field $\boldsymbol{E}$ is a real value.

If a plane wave propagates in an arbitrary direction we can define a vector $\boldsymbol{k} = \{k_x, k_y, k_z\}$.which points into the direction of the wave propagation and which is named **wave vector** with the amount

$$|\boldsymbol{k}| = k = \frac{2\pi}{\lambda}. \tag{7.10b}$$

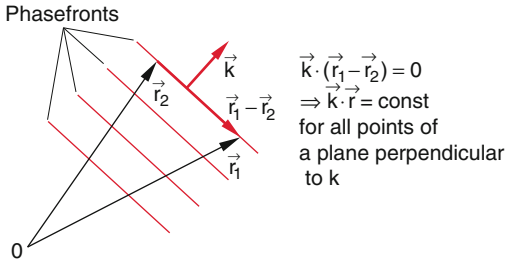


Fig. 7.3 Plane wave in the propagation direction of the wave-vector $\boldsymbol{k}$. The phase planes are the planes $\boldsymbol{k} \cdot \boldsymbol{r}$ = const., perpendicular to the vector $\boldsymbol{k}$

The phase surfaces are planes perpendicular to $\boldsymbol{k}$. The wave vector $\boldsymbol{k}$ is therefore a normal vector on the phase planes, (Fig. 7.3). The complex representation of such waves is in a shorthand form

$$\boldsymbol{E} = \boldsymbol{A}_0 \cdot \mathrm{e}^{\mathrm{i}(\boldsymbol{k}\cdot\boldsymbol{r}-\omega t)}. \tag{7.11}$$

For $\boldsymbol{k}$= {0, 0, $k_z = k$}

Because of $\boldsymbol{k} \cdot \boldsymbol{r} = k_x x + k_y y + k_z z = kz$ Eq. (7.11) transfers into the form of (7.9d).

7.4 Polarization of Electromagnetic Waves

The polarization of an electromagnetic wave is defined by the direction of the electric vector $\boldsymbol{E}$.

7.4.1 Linear Polarized Waves

If the vector $\boldsymbol{E}_0$ of a wave

$$\boldsymbol{E} = \boldsymbol{E}_0 \cdot \cos(\omega t - kz)$$

always points into the same direction perpendicular to $\perp \hat{\boldsymbol{e}}_z$, i.e.

$$\boldsymbol{E}_0 = E_{0x}\hat{\boldsymbol{e}}_x + E_{0y}\hat{\boldsymbol{e}}_y, \tag{7.12}$$

then we call the wave linear polarized (Fig. 7.4). Both components of the wave

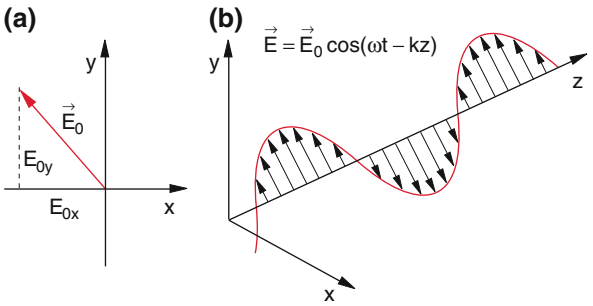


Fig. 7.4 Momentary state of a linearly polarized plane wave $\boldsymbol{E} = \boldsymbol{E}_0 \cdot \cos(\omega t - kz)$. **a**) Direction of the vector $\boldsymbol{E}$ in the x-y plane. **b**) Representation of the electric field vector $\boldsymbol{E}(z, t = t_1)$

$$E_x = E_{0x} \cdot \cos(\omega t - kz)$$
$$E_y = E_{0y} \cdot \cos(\omega t - kz)$$

oscillate in phase (Fig. 7.4a).

7.4.2 Circular Polarization

If the amounts E_{0x} and E_{0y} are equal ($E_{0x} = E_{0y} = E_0$) but the corresponding phases differ by 90° the wave is described by

$$\begin{aligned} E_x &= E_{0x} \cdot \cos(\omega t - kz) \\ E_y &= E_{0y} \cdot \cos\left(\omega t - kz - \frac{\pi}{2}\right) \\ &= E_{0y} \cdot \sin(\omega t - kz) \end{aligned} \qquad (7.13a)$$

The arrow head of the vector

$$\begin{aligned} \boldsymbol{E}(z=0,t) &= E_x \hat{\boldsymbol{e}}_x + E_y \hat{\boldsymbol{e}}_y \\ &= E_0(\hat{\boldsymbol{e}}_x \cos(\omega t)) + \hat{\boldsymbol{e}}_y \sin(\omega \cdot t)) \end{aligned}$$

moves on a circle in the xy-plane with the angular frequency $\omega = \mathrm{d}\varphi/\mathrm{d}t$, i.e. $\varphi = \omega \cdot t$. The electric field vector $\boldsymbol{E}$ with its amount $|\boldsymbol{E}| = E_0$ describes a circular helix about the z-axis (Fig. 7.5).

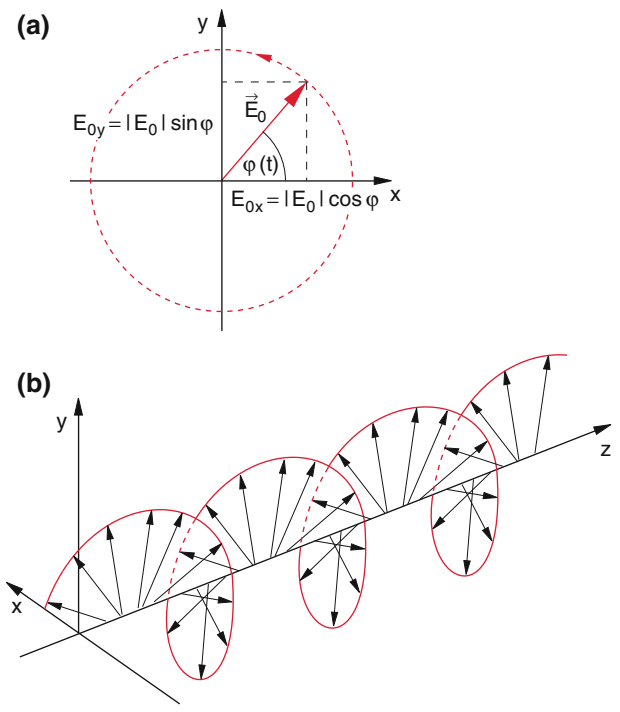


Fig. 7.5 Left-circular polarized electromagnetic wave (σ^+-wave). **a**) $E_0(x, y)$, the z-axis is directed into the page. **b**) Three-dimensional representation

As component representation we can write (7.13a) in a complex notation (Sect. 7.3)

$$\begin{aligned} E_x &= A_0 \cdot \mathrm{e}^{\mathrm{i}(\omega t - kz)} = A_0[\cos(\omega t - kz) + \mathrm{i}\sin(\omega t - kz)] \\ E_y &= A_0 \cdot \mathrm{i} \cdot \mathrm{e}^{\mathrm{i}(\omega t - kz)} = A_0[-\sin(\omega t - kz) + \mathrm{i}\cos(\omega t - kz)] \\ &\text{with } A_0 = \frac{1}{2}E_0. \end{aligned} \qquad (7.13b)$$

Comment

Equation (7.13b) describes light, where the end of the $\boldsymbol{E}$-vector moves on a right-hand screw when the observer looks in the direction of propagation. We will denote it as σ^+-light. In the older literature it is called left-circular polarized, because $\boldsymbol{E}$ runs through a left-handed screw, if looking into the opposite direction of propagation, i.e. into the light source (Fig. 7.5a). We have the following assignments:

$$\sigma^- \rightarrow \text{right circular polarized;}$$
$$\sigma^+ \rightarrow \text{left circular polarized}$$

The new notation σ^+ and σ^- is more meaningful, because the angular momentum $\hbar \cdot \boldsymbol{k}/|\boldsymbol{k}|$ of a circular polarized wave for σ^+-light points in the direction of propagation $\boldsymbol{k}$, for σ^- light into the opposite -$\boldsymbol{k}$-direction (see Sect. 9.6.7).

7.4.3 Elliptical Polarized Waves

If $E_{0x} \neq E_{0y}$ or if the phase difference $\Delta\varphi$ between the components E_{0x} and E_{0y} are not exactly $\pi/2$, then the end of the vector $\boldsymbol{E}$ moves on an elliptic spiral. Such waves are called elliptical polarized.

7.4.4 Unpolarized Waves

If the direction of the vector $\boldsymbol{E}_0$ of the wave (7.9a–7.9e) changes statistically, we call such a wave unpolarized. Light waves are generally unpolarized, because they are a superposition of the emission from many atomic dipoles in randomly distributed directions with statistical phases

In the next chapter we will explain the generation and measurement of polarized light.

7.5 The Magnetic Field of Electromagnetic Waves

Applying the differential operator *curl* to a wave $\boldsymbol{E} = E_0 \cdot \boldsymbol{e}_x \cdot \mathrm{e}^{\mathrm{i}(\omega t - kz)}$ that is linear polarized in the x-direction we obtain

$$(\nabla \times \boldsymbol{E})_x = 0; \quad (\nabla \times \boldsymbol{E})_z = 0;$$
$$(\nabla \times \boldsymbol{E})_y = \frac{\partial E_x}{\partial z}. \tag{7.14}$$

From the Maxwell equation

$$\frac{\partial \boldsymbol{B}}{\partial t} = -(\nabla \times \boldsymbol{E}) \tag{7.15a}$$

We therefore get

$$\frac{\partial B_x}{\partial t} = \frac{\partial B_z}{\partial t} = 0 \tag{7.15b}$$

Which yields $B_x(t)$ = constant and $B_z(t)$ = constant.

The solutions for the B_x- and the B_z-component give only time independent fields that do not contribute to the real wave. We can choose these boundary conditions in a way that the constants become zero. The $\boldsymbol{B}$-field then has only a y-component. From (7.14) follows

$$-\frac{\partial B_y}{\partial t} = \frac{\partial E_x}{\partial z} = -\mathrm{i}kE_x,$$

Which gives after integration with respect to time

$$B_y = \mathrm{i}kE_0 \int \mathrm{e}^{\mathrm{i}(\omega t - kz)}\mathrm{d}t$$
$$= \frac{k}{\omega} E_0 \mathrm{e}^{\mathrm{i}(\omega t - kz)}. \tag{7.16}$$

With $\boldsymbol{E} = \{E_x, 0, 0\}$ and $\boldsymbol{B} = \{0, B_y, 0\}$ we see that $\boldsymbol{E}$ and $\boldsymbol{B}$ are orthogonal (Fig. 7.6). Both vectors are also orthogonal to the propagation vector $\boldsymbol{k}$. We describe this by the vector equation

$$\boldsymbol{B} = \frac{1}{\omega}(\boldsymbol{k} \times \boldsymbol{E}) \tag{7.16a}$$

With the relation $\omega/k = c$ we get $|\boldsymbol{B}| = |\boldsymbol{E}|/c$.

The electric and magnetic field vectors of a plane electromagnetic wave are orthogonal to each other and to the propagation vector $\boldsymbol{k}$. Both fields oscillate in phase. The amount of $\boldsymbol{B}$ is

$$|\boldsymbol{B}| = \frac{1}{c}|\boldsymbol{E}|. \tag{7.17}$$

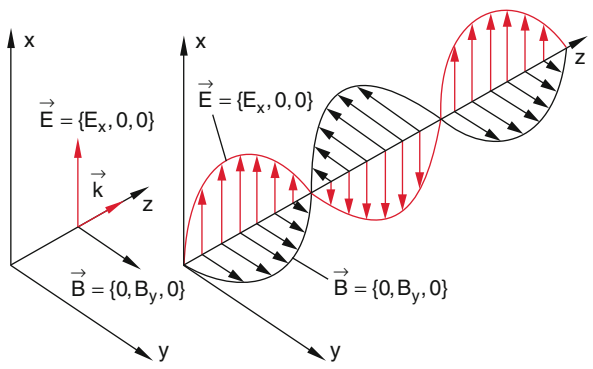


Fig. 7.6 Electric and magnetic field vectors of a linearly polarized plane wave

Examples

1. A 100 W light bulb emits in the range of visible light a power of 5 W. In a distance of 2 m an area of 0.1 m^2 receive the radiation power of 0.1 W. Then the electric field at this point is $|\boldsymbol{E}|$ = 6 V/m, but the magnetic field is only $|B| = |E|/c = 2 \times 10^{-8}$ Vs/m^2.
2. At a wavelength of λ = 500 nm we filter from the radiation of the sun a spectral interval of $\Delta\lambda$ = 1 nm. Then the transmitted green light has an intensity of about 4 W/m^2 at the surface of the earth. This results in an electric field of about 40 V/m. The amount of the magnetic field is then B = 3.3 × 10^{-9} × 40 Vs/m^2 = 1.3 × 10^{-7} T = 1.3 × 10^{-3} G (Gauss). This is much smaller than the mean magnetic field of the earth of 0.2 G.

The cause for the effect of light onto matter (for example exposure of a photographic film, stimulation of the retina cells in our eyes) is mainly due to the electric part of the wave. The magnetic part has, especially in the visible range of the spectrum, a minor influence.

Note

- Only at far distances from the Hertzian dipole ($r \gg d_0$) are the electric field $\boldsymbol{E}(t)$ and the magnetic field $\boldsymbol{B}(t)$ in phase. Close to the dipole the first term in (6.30) for the magnetic field dominates. This term is proportional to dp/dt and therefore has another phase than the second term, which is proportional to d^2p/dt^2. For the electric field E the first term in (6.34) dominates which depends on dp/dt and on d^2p/dt^2. Directly at the dipole the phases of $\boldsymbol{B}$ and $\boldsymbol{E}$ differ by 90°. This can be seen from Fig. 6.2 for current and voltage of an oscillator and also from the graphs of magnetic and electric field lines in Figs. 6.24 and 6.25. In the transition region between near-field and far-field the phases change continuously until $\boldsymbol{E}$ and $\boldsymbol{B}$ have the same phase.
- The relations $\boldsymbol{B} \perp \boldsymbol{E}$ and $\mathbf{E}, \mathbf{B}, \perp \mathrm{k}$ (transversality of electromagnetic waves) is generally valid only in vacuum. In case of currents or volume charge densities, **B** and **E** need not be perpendicular.

7.6 Transport of Energy and Momentum by Electromagnetic Waves

In Sect. 4.4 we have derived for the energy density of electromagnetic fields the expression

$$w_{\text{em}} = \frac{1}{2}\varepsilon_0(E^2 + c^2B^2) = \varepsilon_0 E^2 \tag{7.18}$$

with $B^2 = E^2/c^2$.

This energy density of an electromagnetic wave is transported with the propagation speed c in the direction of the wave vector $\boldsymbol{k}$ (Fig. 7.8). We denote the amount of energy that is transported within the unit of time through a unit area perpendicular to $\boldsymbol{k}$ as **the intensity** $\boldsymbol{I}$ or energy flux.

$$I = c \cdot \varepsilon_0 \cdot E^2. \tag{7.19}$$

Since $\boldsymbol{E} = \boldsymbol{E}_0 \cos(\omega t - \boldsymbol{k} \cdot \boldsymbol{r})$ is a periodic function in time the intensity of a linear plane wave varies periodically with the frequency 2ω (because $\cos^2\omega t = (1 + \cos 2\omega t)$).

$$I(t) = I_0 \cdot \cos^2(\omega t - \boldsymbol{k} \cdot \boldsymbol{r}) \quad \text{mit} \quad I_0 = c\varepsilon_0 E_0^2$$

The intensity is two times per period T equal to zero. The mean value of I is with $\langle \cos^2\omega t \rangle = ½$

$$\langle I(t) \rangle = \frac{1}{2} c \cdot \varepsilon_0 E_0^2. \tag{7.20a}$$

Example

The intensity of the sun radiation outside of our atmosphere is 1.2 kW/m^2 (solar constant). The total power radiated by the sun to the earth is then P = 1.2 kW · $2\pi R^2 \approx 3 \times 10^{14}$ W with R = 6.7 × 10^6 m = radius of the earth. Due to absorption and scattering in the atmosphere the intensity reaching the surface of the earth is only about 700 W/m^2 A collector surface of 100 m^2 (Fig. 7.7) can then collect at most a power of 70 kW. With an efficiency of 20% this gives about 14 kW at noon time. On the average for a sunny day with 10 h sunshine at a latitude of 45° the electric output energy of about 10 kWh (because in the morning and evening the sun radiation power is much lower. In summary: Although the sun radiation to the whole earth yields a power which is by far the highest energy supply available, the power getting from small areas of sun radiation collectors is modest and is not sufficient to supply the demand.

Fig. 7.7 Collector field for sun radiation (http://www.yaacool-bio.de/index.php?article=1637)

Comment

1. Using the complex

Notation $\boldsymbol{E} = \boldsymbol{A}_0 \cdot e^{i(\omega t - \boldsymbol{k} \cdot \boldsymbol{r})} + c.c$ (see Sect. 7.3) and $I = c \cdot \varepsilon_0 \cdot |E|^2 = 4c \cdot \varepsilon_0 |A_0|^2 \cdot \cos^2(\omega t - \boldsymbol{kr})$ the mean value becomes

$$\langle I(t) \rangle = 2c\varepsilon_0 |A_0|^2. \tag{7.20b}$$

2. Circular polarized waves show a phase difference of 90° between E_x- and E_y- component. Therefore the intensity

$$\begin{aligned} I &= c\varepsilon_0(E_x^2 + E_y^2) \\ &= c\varepsilon_0 E_0^2[\sin^2(\omega t - \boldsymbol{k} \cdot \boldsymbol{r}) + \cos^2(\omega t - \boldsymbol{k} \cdot \boldsymbol{r})] \\ &= c\varepsilon_0 E_0^2 \end{aligned} \tag{7.20c}$$

is constant in time and never becomes zero (contrary to the linear polarized wave).

The direction of the energy flux is defined by the **Poynting vector**

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} \tag{7.21a}$$

In vacuum it is with $c^2 = 1/(\mu_0\varepsilon_0)$

$$\boldsymbol{S} = \varepsilon_0 \cdot c^2(\boldsymbol{E} \times \boldsymbol{B}) \tag{7.21b}$$

The amount of $\boldsymbol{S}$ is with (7.17) and (7.20a–7.20c)

$$\begin{aligned} S = |\boldsymbol{S}| &= \varepsilon_0 c^2 |\boldsymbol{E}| \cdot |\boldsymbol{B}| \\ &= \varepsilon_0 c E^2 = I, \end{aligned} \tag{7.22}$$

equals to the intensity I of the wave. The unit for S is

$$[S] = \text{W/m}^2.$$

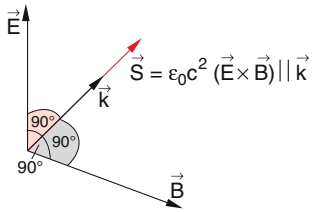


Fig. 7.8 Energy transport by a plane wave into the direction of the pointing vector **S**

For a plane electromagnetic wave in vacuum is $\boldsymbol{E} \perp \boldsymbol{B}$;

$$\boldsymbol{E} \perp \boldsymbol{k}, \text{ and } \boldsymbol{B} \perp \boldsymbol{k}.$$

Then the Poynting vector $\boldsymbol{S} = \varepsilon_0 c^2 (\boldsymbol{E} \times \boldsymbol{B})$ must point into the direction of propagation $\boldsymbol{k}$ of the wave (Fig. 7.8). We can visualize Eq. (7.22) in the following way.

We consider a volume V in vacuum (Fig. 7.9) which contains the field energy

$$W_{\text{em}} = \int \varepsilon_0 E^2 \cdot \mathrm{d}V$$

The energy flow per unit time through the surface A of this volume V must be equal to the rate with which the energy included in V decreases.

$$-\frac{\partial}{\partial t} \int \varepsilon_0 E^2 \cdot \mathrm{d}V = \oint \boldsymbol{S} \cdot \mathrm{d}\boldsymbol{A} = \int_V \text{div } \boldsymbol{S} \cdot \mathrm{d}V, \quad (7.23a)$$

The last equation in (7.23a) is Gauss' theorem. Because it is valid for any volume (conservation of energy) it follows for the integrand

$$-\frac{\partial}{\partial t}(\varepsilon_0 E^2) = \text{div } \boldsymbol{S}. \quad (7.23b)$$

Since **div S** describes the productivity of the source of the electromagnetic field, i.e. the amount of energy per unit time and per unit volume that flows out (or into) the whole volume, it follows from (7.23a) that $\boldsymbol{S}$ is the field energy flux through the surface that encloses the volume V (Fig. 7.9). The amount $|\boldsymbol{S}| = S$ is according to (7.18) the intensity I of the electromagnetic wave that leaves the volume V.

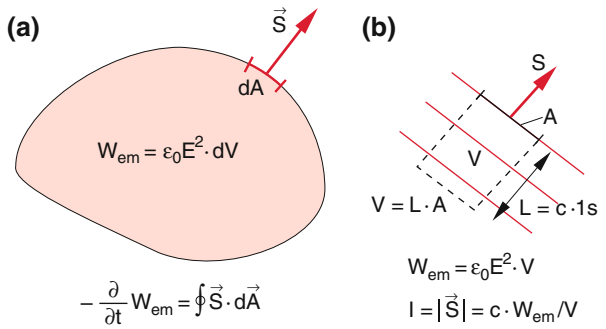


Fig. 7.9 **a**) Illustration of the pointing vector as vector of energy flux per unit area. **b**) For a plane wave **S** is perpendicular to the phase planes

Note This is not valid in anisotropic media, where the direction of the Poynting vector $\boldsymbol{S}$ (energy flux) and the vector $\boldsymbol{k}$ (direction of propagation) do not point into the same direction.

Examples

1. While a capacitor is charged, between its plates an electric field $\boldsymbol{E}$ develops and an electric current $I = \mathrm{d}Q/\mathrm{d}t$. flows through the connecting wires. Around the increasing electric field in the volume between the plates of the capacitor a circular magnetic field $\boldsymbol{B}$ is formed (Fig. 7.10a).
   The Poynting vector $\boldsymbol{S} = \varepsilon_0 \cdot c^2 (\boldsymbol{E} \times \mathrm{B})$ points radially to the center. This implies that the energy flux which builds up the electric field is not parallel to the supplying wires in z-direction, as one would suppose, but the energy flows radially from outside into the field.
2. Through a straight wire with the resistance R flows a constant current I, creating Joule's heat $\mathrm{d}W_{\text{el}}/\mathrm{d}t = I^2 R$. Of course for stationary equilibrium the dissipated power must be supplied from external sources. Also in this case the Poynting vector is directed radially from outside into the wire but not along the wire (Fig. 7.10b). The explanation is as follows. The current carying electrons move with the very small drift velocity v_D (see Sect. 2.2). At a current of 10 A through a wire of 1 mm^2 cross section is the drift velocity $v_{\text{D}} = 8$ mm/s. The electric and magnetic fields created by the current, propagate with the speed of light $v = (\varepsilon \cdot \varepsilon_0 \cdot \mu \cdot \mu_0)^{-1/2}$ along the wire, Therefore the energy is transported by the electromagnetic field but not by the physical carriers of charge.

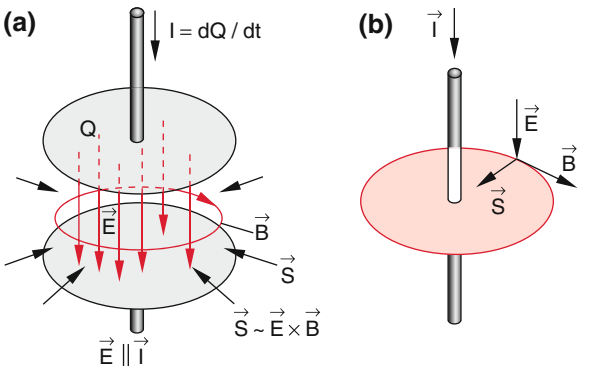


Fig. 7.10 Direction of the pointing vector **a**) during charging of a capacitor, **b**) for the supply of the energy, spent by Joule's heating in a wire carrying the electric current I

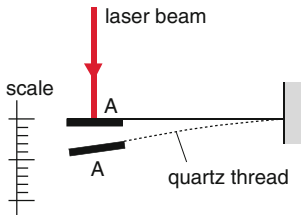


Fig. 7.11 Measurement of radiation pressure through the deflection of a sensitive quartz balance with absorbing surface with area A

A plane electromagnetic wave causes not only an energy flow $\boldsymbol{S}$ but also a momentum transfer per unit volume

$$\boldsymbol{\pi}_{\mathrm{St}} = \frac{1}{c^2}\boldsymbol{S} = \varepsilon_0(\boldsymbol{E} \times \boldsymbol{B}). \tag{7.24}$$

The momentum has the direction of the Poynting vector $\boldsymbol{S}$ and the amount

$$|\boldsymbol{\pi}_{\mathrm{St}}| = \varepsilon_0 \cdot E \cdot B = w_{\mathrm{em}}/c = I/c^2, \tag{7.25}$$

where I is the intensity of the wave.

The momentum density of the electromagnetic wave is then $\pi_{\mathrm{St}} = w_{\mathrm{em}}/c$. A particle which would move with nearly the speed of light c had the energy $E = mc^2$ and the momentum $p = mc = E/c$ (see Vol. 1, Sect. 4.4). Therefore we can in an analogue way ascribe to the electromagnetic wave a mass density

$$\rho_{\mathrm{m}} = w_{\mathrm{em}}/c^2 = \left(\varepsilon_0/c^2\right)E^2.$$

If an electromagnetic wave is absorbed by a particle (see Sect. 8.2) its momentum is transferred to this particle which therefore suffers a repulsion. If the wave is reflected by the particle, twice the momentum is transferred. The transferred momentum per unit time and area corresponds to the pressure onto the surface.

Therefore the radiation pressure of a plane wave perpendicular to a completely absorbing surface of a body is

$$p_{\mathrm{St}} = c \cdot |\boldsymbol{\pi}_{\mathrm{St}}| = \varepsilon_0 E^2 = w_{\mathrm{em}} \tag{7.26}$$

where the unit of energy density of the electromagnetic wave is equal to the unit of pressure

$$[w_{\mathrm{em}}] = 1\,\frac{\mathrm{Ws}}{\mathrm{m}^3} = 1\,\frac{\mathrm{N}}{\mathrm{m}^2}$$

The radiation pressure can be measured by a very sensitive balance (Fig. 7.11).

Examples

1. A light beam with the mean power $\bar{P}_{\mathrm{el}} = 10\,\mathrm{W}$ falls normally on an absorbing area $A = 1\,\mathrm{mm}^2$. It transfers per unit time the momentum $|\mathrm{d}p/\mathrm{d}t| = \pi_{\mathrm{St}} \cdot A \cdot c$ onto the area. The amount of the repulsive force is

$$|\boldsymbol{F}| = \frac{\mathrm{d}p}{\mathrm{d}t} = \bar{P}_{\mathrm{el}}/c. \tag{7.27}$$

Its amount is $F = 3.3 \times 10^{-8}$ N. The radiation pressure $p_{\mathrm{St}} = F/A = 3.3 \times 10^{-2}$ Pa is very low and can only be measured for high light powers and by sensitive balances.

2. With a pulsed high power laser with intensities up to 10^{18} W/cm^2 radiation pressure of 10^9 bar = 10^{14} Pa can be realized.
3. With nearly frictionless bearings in vacuum and with the help of radiation pressure one can operate a light mill [1]. It consists of four wings that are reflecting light at one side and absorbing it at the other side. The wings are mounted so that they can rotate about a common vertical axis. The transferred momentum is at the reflecting side twice that at the absorbing area. This results in a net angular momentum that rotates the wings against the low friction of its bearings. The commercially available light mills rotate in the opposite direction as that described above (Fig. 7.12). What is the difference?
4. Hint: There is no vacuum inside the mill (see Example 7.10).
5. The radiation pressure of the sun is one of the reasons for the curvature of the tails of comets. The tail of comets consists of matter from the core of the comet that evaporates while the comet moves near the sun. The radiation of the sun supplies the necessary energy. The tail consists of neutral molecules, ions and dust. The electrically charged ions are deflected by the magnetic field of the sun. The dust-particles are more affected by the radiation pressure. Therefore one observes two tails with different curvature (Fig. 7.13) where the dust-tail has the stronger curvature [2, 3].

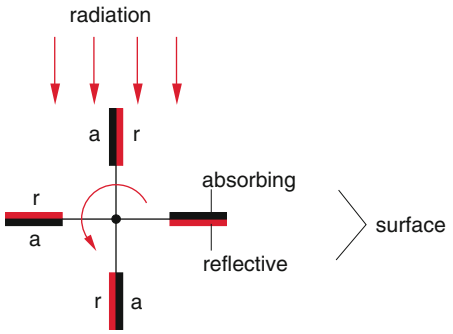


Fig. 7.12 Light mill in a vacuum container. For the given design it rotates anticlockwise

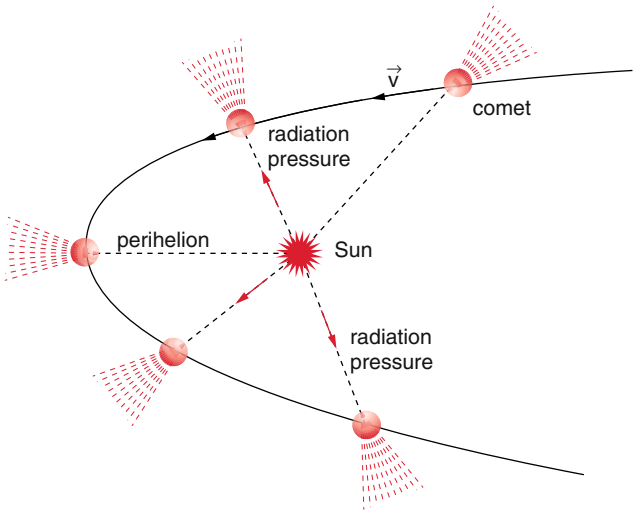


Fig. 7.13 Deflection of the comet tail by the radiation pressure from the sun

Note The solar wind (protons and electrons) causes inhomogeneous electric and magnetic fields and divide the comet's tail sometimes in more than two components.

7.7 Measurement of the Speed of Light

According to our present knowledge the speed of light is in vacuum independent of its frequency ω. That means, that phase- and group-velocity in vacuum are always equal; there is no dispersion! (see Vol. 1, Sect. 11.9.7).

$$v_{\mathrm{Ph}} = v_{\mathrm{G}} = \frac{\omega}{k} = c. \tag{7.28}$$

The speed of light can, therefore, be measured at any frequency. Up to now most of the measurements were done with visible light. That is the reason for naming c the speed of light, although the value is valid for all electromagnetic waves of the complete spectrum.

7.7.1 The Astronomical Method of Ole Roemer

The oldest method to determine the speed of light is based on astronomical observations. Many astronomers have measured the orbital period of the moons of Jupiter with high precision, because the time of darkening—the moons are concealed by Jupiter—and the time of reappearance could be observed very well. *Ole Roemer* (1644–1710) found out, that the available tables reproduce the period of revolution well, if the earth is near to Jupiter (position 1 in Fig. 7.14)—Jupiter in opposition to the sun—but the observed darkenings were 22 min later, if Jupiter where in conjunction (position 2 of the earth).

Contrary to other scholars of his time, Roemer traced back the results of the observations to the different times the light

Fig. 7.14 Photograph of the comet Mrkos 1957d where the tail is split (with kind permission of the Hale observatory)

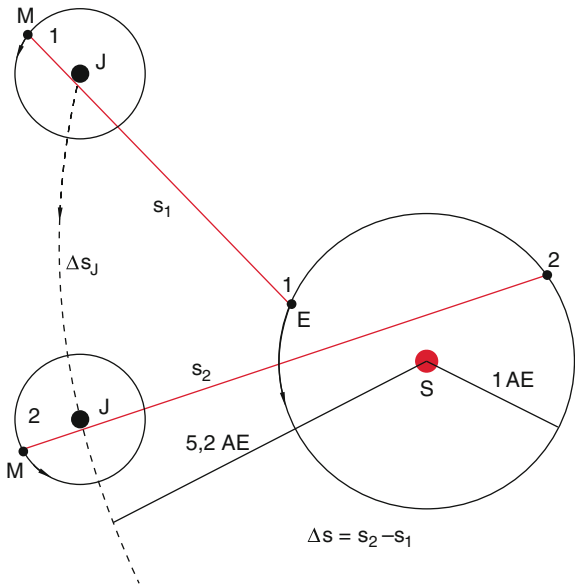


Fig. 7.15 Determination of the speed of light by the astronomical method of Ole Roemer. The drawing is not to scale

needed between Jupiter and earth for the two positions 1 and 2 in Fig. 7.15 with the path difference $\Delta s = s_2 - s_1$. Roemer could show, that the speed of light has a finite amount and is

not infinitely large, contrary to the opinion of Descartes. For the exact determination of the path difference, one has to take into account, that Jupiter has moved by the arc length Δs_j during the time $\Delta t = t_2 - t_1$. The diameter D of the earth orbit has been well known ($D \approx 3 \times 10^{11}$m). So, Roemer could calculate the speed of light from the difference of his measurements and the data in the available tables.

But Roemer did not publish a definite value of c, maybe because he believed his measurements to be not exact enough [4]. Later Huygens published a value of c between 220,000 and 300,000 km/s [5], a value, that includes the true value.

7.7.2 Cogwheel Method by Fizeau

While Roemer used a very large distance (3×10^{11} m) *Armand Fizeau* (1819–1896) had improved the time measurements so much that he could use a distance on earth to determine the speed of light. He used an experimental setup according to Fig. 7.16. An astronomical telescope collimated the light coming from an extended source LQ to a parallel beam that was reflected by the mirror S at a distance d. A part of the reflected light was split by a beam splitter BS and the transmitted light reached the observer.

A fast rotating cogwheel CW in the focal plane of the lens L_1 periodically interrupts the-light beam, so light pulses of duration T_1 and frequency $\nu = 1/\Delta T = 1/(2T_1)$ are emitted if tooth and gap of the cogwheel have the same width.

If the cogwheel rotates with such an angular velocity ω that the light pulse transmitted by the gap n returns back at the next gap $(n + 1)$ the observer sees light.,

At a faster rotation of the cogwheel the reflected beam meets a tooth and one observes darkness. At twice the rotational velocity, 2ω, the reflected light pulse meets again a gap $n + 2$, and so on.

Assume the cogwheel has N teeth and rotates with ω, then the time between two successive gaps is

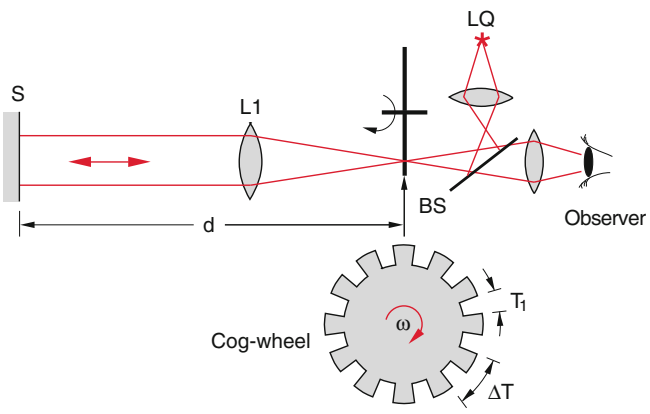


Fig. 7.16 Measurement of the light velocity with the cog-wheel method of Fizeau

$$\Delta T = \frac{2\pi}{\omega}\frac{1}{N},$$

The speed of light is then calculated as

$$c = \frac{2d}{\Delta T} = \frac{d \cdot N \cdot \omega}{\pi} = 2dN \cdot f$$

where $f = \omega/2\pi$ is the rotation frequency.

Fizeau used a distance of d = 8.6 km between the summits of two mountains. His cogwheel had N = 720 teeth and rotated with the frequency f = 25.2 Hz. The light was interrupted at the frequency $\nu = Nf = 720 \times 25.2$ Hz. His result was c = 315,000 km/s. The difference of 5% to the accepted value today results mainly from errors in determining the rotation frequency [6].

7.7.3 The Rotating Mirror of Foucault

A much higher accuracy was achieved by *Bernard Leon Foucault* in 1850 with his method of the rotating mirror (Fig. 7.17). A lens L focusses the light source onto the aperture B and after a distance L the light is reflected by a rotating plane mirror M_1 onto a fixed concave mirror M_2. Without rotation of M_1 the light reflected by M_2 would hit the aperture B again. When M_1 rotates with the frequency ω the reflected light produces on a photographic plate an image of the slit S which is shifted against the slit S by the distance Δx.

The mirror M_1 with a rotation period $T = 2\pi/\omega$ rotates during the time

$$\Delta T = \frac{2d}{c} = T \cdot \frac{\alpha}{2\pi} = \frac{\alpha}{\omega}. \qquad \text{(a)}$$

by the angle $\alpha = 2d \cdot \omega/c$. The distance Δx is then

$$\Delta x = L \cdot \tan 2\alpha \approx 2L \cdot \alpha. \qquad \text{(b)}$$

The velocity of light is now obtained from (a) and (b) as

$$c = 2d \cdot \omega/\alpha = 4d \cdot \omega \cdot L/\Delta x \qquad \text{(c)}$$

The measurement of the speed of light c by this method is reduced to the measurement of the distances d, L and Δx and the rotation frequency ω.

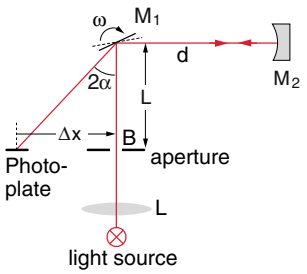


Fig. 7.17 Principle of the rotating mirror method of Foucault

Repeating Foucault's experiment today, we use a laser as the light source that can be much better collimated. The rotation frequency of the plane mirror is driven by an electric motor and can be determined easily and very accurately. Therefore much smaller distances d are sufficient to reach an acceptable accuracy. The experiment can then even be used as demonstration in the lecture hall [7].

7.7.4 Phase Method

Instead of a cogwheel as in Fizeau's experiment today optical modulators are available to interrupt the light at much higher frequencies f (Fig. 7.18). With a He–Ne-laser that produces a collimated parallel beam of light which passes through a Pockels cell. A Pockels cell is an optical modulator that rotates the polarization plane following the high frequency of an applied high voltage. Together with a polarizer it acts as an intensity modulator

The transmitted intensity I_t behind the polarizer P is modulated at the frequency f according to

$$I_t = \frac{1}{2} I_0 [1 + \cos^2(2\pi f t)] \qquad \text{(d)}$$

Part of the transmitted beam is split by the beam splitter BS onto the fast photo detector PD_1. After reflection at a retroreflector the other part of the light is imaged onto the photo detector PD_2. The phase shift

$$\Delta\varphi = \Delta T \cdot 2\pi f = (s_2 - s_1) \cdot 2\pi f / c \qquad \text{(e)}$$

between the two modulated laser beams is measured and yields the speed c = ($\Delta s \cdot 2\pi f / \Delta\varphi$ of light.

Example

$f = 10^7 \text{Hz}, \Delta s = 3\text{m}, \Rightarrow \Delta\varphi = 2\pi f \cdot \Delta T = 2\pi f \cdot 2d/c \approx 72°$. The phase φ can be measured to 0.1°. Therefore the result is accurate within ±0.14%.

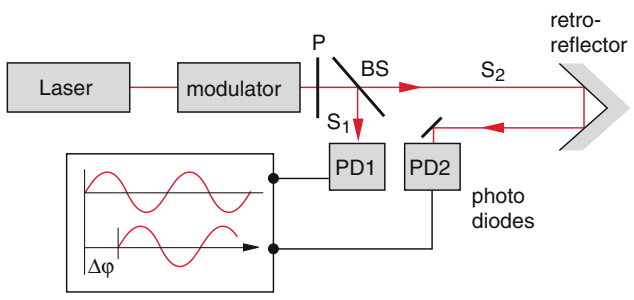


Fig. 7.18 Measuring the velocity with the phase method. P = polarizer, BS = beamsplitter

7.7.5 Determination of *c* by Measurements of Frequency and Wavelength

From the relation

$$c = \nu \cdot \lambda$$

for electromagnetic waves the velocity of light c can be determined, if both the wavelength λ and the frequency ν can be measured simultaneously. The wavelength λ can be measured with high accuracy using modern interferometric techniques [8]. Optical frequencies can be measured only recently, since techniques for division of frequencies and the frequency comb have been developed [9]. The most accurate measured value of the speed of light is obtained by a weighted average of several measurements. The today accepted value is

$$c = 2,99792458 \times 10^8 \text{m/s}.$$

This value is now used to define the unit of length . The new definition of the meter is:

> 1 m is the length, that is travelled by light in vacuum within 1/299792458 s. The speed of light is no longer a quantity that can change by new measurements but has the fixed defined value (see Vol. 1, Sect. 1.6.1).

$$\lambda = c/\nu$$

With this definition only the frequency has to be measured. This is nowadays possible with a much higher accuracy than that of wavelength measurements [10].

Table 7.1 lists some historical measurements of the speed of light and their uncertainties.

Table 7.1 Historical measurements of the speed of light

Year	Author	Method	Measured value given in km/s
1677	Ole Rømer	astronomical	finite, no value given
1678	Huygens	Analysis of Romers measurements	220–300 × 10^3
1849	A. Fizeau	cogwheel method	315 000
1862	L. Foucault	rotating mirror method	298 000
1879	A. Michelson	improved rotating mirror technique	299 910
1926	A. Michelson	interferometer	299 791
1950	L. Essen	Microwave cavity	299 792,5
1973	K. Evenson	measurement of wavelength and frequency of a laser transition	299 792,45
seit 1983	–	todays defined fixed value	299 792,458

7.8 Standing Electromagnetic Waves

Standing electromagnetic waves can be created exactly in the same way as mechanical waves by superposition of travelling waves from opposite directions which have equal frequencies.

7.8.1 Standing Waves in One Direction

Standing waves in one direction arise from reflection of a plane wave that is incident normally onto the plane boundary of a conducting medium (see Vol. 1, Sect. 11.12).

Now we consider a linearly polarized plane wave

$$E = E_{0x}\cos(\omega t - kz)$$

Travelling into the +z-direction, with the electric vector $\boldsymbol{E} = \{E_x,\ 0,\ 0\}$ and the magnetic vector $\boldsymbol{B} = \{0,\ B_y,\ 0\}$ (Fig. 7.19).

Because no tangent component E_x can exist at the surface of a perfect conductor at $z = 0$, we get at the plane $z = 0$

$$\begin{aligned} \boldsymbol{E}(z=0) &= \boldsymbol{E}_{0\mathrm{i}} + \boldsymbol{E}_{0\mathrm{r}} = \boldsymbol{0} \\ &\Rightarrow \boldsymbol{E}_{0\mathrm{i}} = -\boldsymbol{E}_{0\mathrm{r}}. \end{aligned} \tag{7.29}$$

The superposition of incident wave $\boldsymbol{E}_\mathrm{i}$ and reflected wave $\boldsymbol{E}_\mathrm{r}$ gives

$$\begin{aligned} \boldsymbol{E}(z,t) &= \boldsymbol{E}_{0\mathrm{i}}\cos(\omega t - kz) + \boldsymbol{E}_{0\mathrm{r}}\cos(\omega t + kz) \\ &= 2\boldsymbol{E}_0 \cdot \sin(kz) \cdot \sin(\omega t) \end{aligned} \tag{7.30}$$

where $\boldsymbol{E}_0 = \boldsymbol{E}_{0\mathrm{i}} = -\boldsymbol{E}_{0\mathrm{r}}$.

The magnetic part is obtained from the relation

$$\frac{\partial E_x}{\partial z} = -\frac{\partial B_y}{\partial t},$$

that follows from the Maxwell equation $\mathbf{rot}\,\boldsymbol{E} = -\dot{\boldsymbol{B}}$

$$\boldsymbol{B}(z,t) = 2\boldsymbol{B}_0\cos(kz)\cdot\cos(\omega t) \tag{7.31}$$

with $\boldsymbol{B}_0 = \{0, (k/\omega)\cdot E_0, 0\}$.

Contrary to travelling waves at large distances from the source where $\boldsymbol{E}$ and $\boldsymbol{B}$ are in phase, we have now between the maxima of $\boldsymbol{E}$ and $\boldsymbol{B}$ a spatial displacement of $\lambda/4$ and a temporal difference of $T/4 = \pi/2\omega$.

The reason for this phase shift is the phase jump of 180° for the electric component $\boldsymbol{E}$ at the reflection (7.29). There is no such phase shift for the magnetic component (see Sect. 8.4) because it has according to (7.31) its maximum at $z = 0$ and suffers no phase reversion at reflection.

Such standing electromagnetic waves with wavelengths in the range of about 0.1–1 m can be demonstrated for a dipole antenna with a sensitive light bulb in its center, which is moved in z-direction (Fig. 7.20). At the maxima of the electric field E the lamp lights up, while at the nodes it is dark.

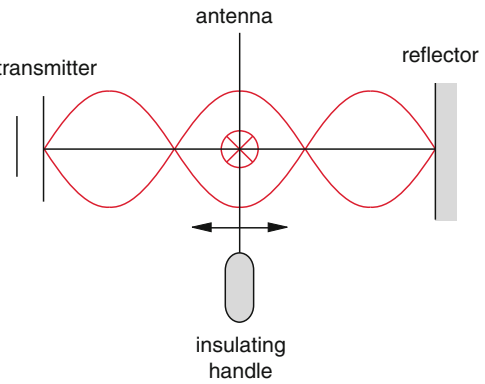


Fig. 7.20 Detection of a one-dimensional electro-magnetic standing wave with a dipole antenna and a light bulb

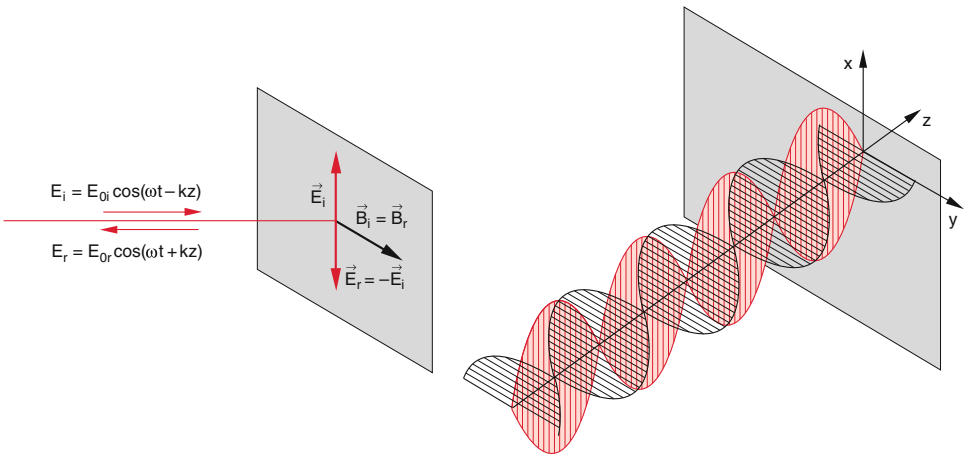


Fig. 7.19 One-dimensional electro-magnetic standing wave generated by superposition of the wave reflected at a conductive plane at $z = 0$ with the incident wave

7.8.2 Three-Dimensional Standing Waves; Cavity Resonators

We consider a cuboid of size a, b, c with perfectly conducting walls (Fig. 7.21). The origin of our Cartesian coordinate system is at one corner and its axes along the edges. So the boundary conditions for the electric field $\boldsymbol{E} = \{E_x, E_y, E_z\}$ demand, that the tangent component at the walls must be zero.

$$\begin{array}{llllll} E_x = 0 & \text{for} & z = 0, c & \text{and} & y = 0, b; \\ E_y = 0 & \text{for} & x = 0, a & \text{and} & z = 0, c; \\ E_z = 0 & \text{for} & x = 0, a & \text{and} & y = 0, b. \end{array} \tag{7.32a}$$

If an electromagnetic wave with wave vector $\boldsymbol{k} = \{k_x, k_y, k_z\}$ is created in a cavity, it is reflected at the walls. The superposition of the various components with wave vectors $\{\pm k_x, \pm k_y, \pm k_z\}$ leads only then to a stationary standing wave, if the boundary conditions

$$k_x = n\pi/a; k_y = m\pi/b; k_z = q\pi/c \tag{7.32b}$$

are fulfilled with integer numbers n, m, q. The amount of the wave vector

$$|\boldsymbol{k}| = \sqrt{k_x^2 + k_y^2 + k_z^2}$$

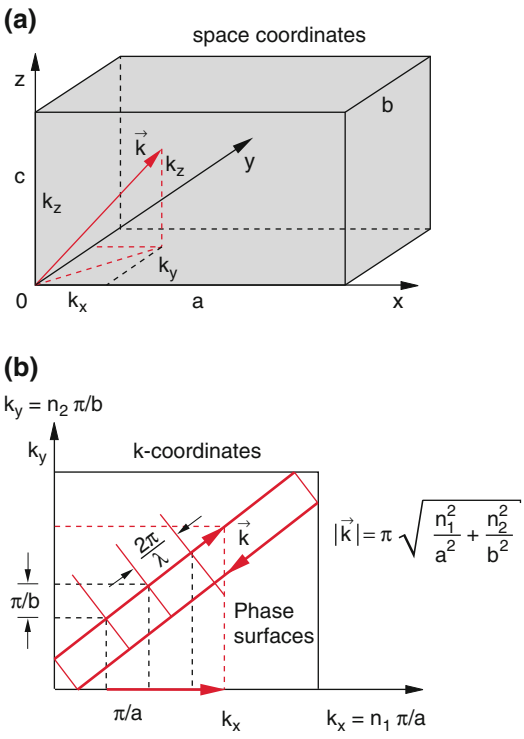


Fig. 7.21 Cuboid made of conductive walls as cavity for standing electro-magnetic waves. **a**) Representation in the spatial domain. **b**) Illustration of the boundary conditions (7.32b) and (7.33) in momentum space

is obtained from the boundary conditions (7.32b) as

$$|\boldsymbol{k}| = k = \pi\sqrt{\frac{n^2}{a^2} + \frac{m^2}{b^2} + \frac{q^2}{c^2}}. \tag{7.33}$$

The possible frequencies ω of an arbitrary standing wave in the cuboid are with $\omega = c_0 \cdot k$ (here c_0 is the speed of light in vacuum)

$$\omega = c_0 \cdot \pi\sqrt{\frac{n^2}{a^2} + \frac{m^2}{b^2} + \frac{q^2}{c^2}}. \tag{7.34}$$

In the cuboid only such standing waves are possible, that have the following form:

$$\boldsymbol{E}_{n,m,q} = \boldsymbol{E}_0(n, m, q) \cdot \cos \omega t$$

With $\boldsymbol{E}_0 = \{E_{0x}, E_{0y}, E_{0z}\}$ and

$$\begin{aligned} E_{0x} &= A \cdot \cos\left(\frac{\pi n}{a}x\right) \sin\left(\frac{\pi m}{b}y\right) \sin\left(\frac{\pi q}{c}z\right), \\ E_{0y} &= B \cdot \sin\left(\frac{\pi n}{a}x\right) \cos\left(\frac{\pi m}{b}y\right) \sin\left(\frac{\pi q}{c}z\right), \\ E_{0z} &= C \cdot \sin\left(\frac{\pi n}{a}x\right) \sin\left(\frac{\pi m}{b}y\right) \cos\left(\frac{\pi q}{c}z\right). \end{aligned} \tag{7.35}$$

Their amplitude $\boldsymbol{E}_0$ is perpendicular to the wave vector $\boldsymbol{k}$ that fulfills the boundary condition (7.32b).

We name the ideal conducting box a *cavity resonator* and the possible standing wave (7.35) its resonant oscillations. Their frequencies are called eigen-frequencies or natural frequencies.

Our next question is, how many frequencies ω up to a given upper limit ω_G are possible inside the cavity.

To simplify the calculation we consider the special case of a cube with $c = b = a$. instead of a cuboid. The conditions for the frequencies (7.34) then become

$$\begin{aligned} \omega &= \frac{c_0 \cdot \pi}{a}\sqrt{n^2 + m^2 + q^2} \\ &\Rightarrow n^2 + m^2 + q^2 = \omega^2 a^2/\left(c_0^2 \pi^2\right). \end{aligned} \tag{7.36}$$

In a coordinate system with the axes k_x, k_y, k_z the points (n, m, q) form a lattice with a lattice constant π/a (Fig. 7.22). There are as many natural oscillations in the cavity as lattice points in the $\boldsymbol{k}$-space. In the $\boldsymbol{k}$-space (7.33) represents the equation of a sphere with the radius $|\boldsymbol{k}| = \pi/a\sqrt{n^2 + m^2 + q^2} = \omega/c_0$.

For $n^2 + m^2 + q^2 \gg 1$ the radius of the sphere is large compared to the lattice constant π/a, i.e. $\lambda \ll 2a$. Then the number N_G of lattice points with $n, m, q > 0$ is with a good approximation equal to the number of unit cells $(\pi/a)^3$ in an octant of the sphere (Fig. 7.23). Its volume in the $\boldsymbol{k}$-space is

$$V_k = \frac{1}{8} \cdot \frac{4\pi}{3} k_G^3 = \frac{\pi}{6}\left(\frac{\omega}{c_0}\right)^3. \tag{7.37a}$$

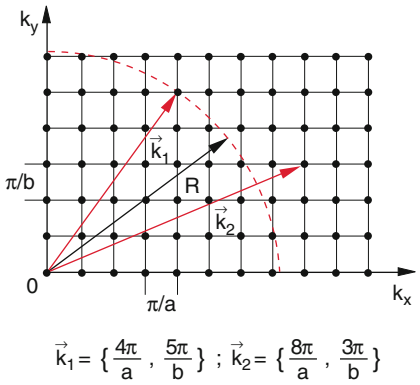


Fig. 7.22 Representation of the k-vectors of possible standing waves in the cavity as grid points in the k-domain

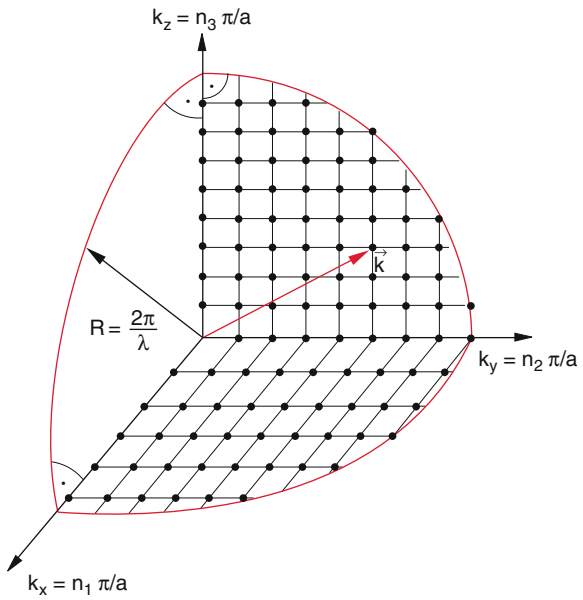


Fig. 7.23 Illustration of the number of possible natural oscillations in a cubic resonator

The number of lattice points is then

$$N_G = V_k/V_E = \frac{\pi}{6}\left(\frac{a\omega}{\pi c_0}\right)^3, \tag{7.37b}$$

where $V_E = (\pi/a)^3$ is the volume of the unit cell in $\boldsymbol{k}$-space.

Now we take into consideration that each standing wave can have an arbitrary direction of its polarization, which can be generated by the linear combination of two orthogonal polarized waves. For a standing wave in z-direction

$$\boldsymbol{E} = \boldsymbol{E}_0 \cdot \sin kz \cdot \sin \omega t$$

is $\boldsymbol{E}_0 = \boldsymbol{E}_{0x}\boldsymbol{e}_x + \boldsymbol{E}_{0y}\boldsymbol{e}_y$),

Then we get the number of possible natural oscillations in the cavity with frequencies ω lower than a given upper limit ω_G

$$N(\omega \le \omega_G) = \frac{\pi}{3}\left(\frac{a \cdot \omega_G}{\pi c_0}\right)^3 = \frac{8\pi v_G^3 a^3}{3c_0^3}, \tag{7.38a}$$

where we have inserted the frequency $v_G = \omega_G/2\pi$.

Dividing by the real volume, $V = a^3$, of the resonator gives the number of modes per unit volume with $v \le v_G$

$$N/V = n = \frac{8\pi v_G^3}{3c_0^3}. \tag{7.38b}$$

Often it is of interest to know the spectral density of modes dn/dv, that is the number of possible natural frequencies per unit volume of the resonator in the frequency interval between v and $v + \Delta v$, with Δv = 1 Hz.

Differentiation of (7.38b) results in

$$\mathrm{d}n/\mathrm{d}v = \frac{8\pi v^2}{c_0^3}, \tag{7.39}$$

where dn/dv is denoted *spectral density of modes*.

Note

- The results above are also obtained in a very general form, when we solve the wave equation

$$\Delta \boldsymbol{E} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}$$

with the boundary conditions $\boldsymbol{E}_t = \boldsymbol{0}$ for $x = 0, a$; $y = 0, b$; $z = 0, c$.

The general stationary solution is the linear combination of the resonator modes (7.35)

$$\boldsymbol{E}(\boldsymbol{r}, t) = \sum_n \sum_m \sum_q \boldsymbol{E}_{n,m,q} \tag{7.40}$$

- If the resonator is not a cuboid, it is not always possible to find analytic solutions. If we have a circular cylinder the solutions are Bessel functions instead of the sinusoidal functions (7.35) for the amplitudes of the resonator modes [11].

7.9 Waves in Wave Guides and Cables

Waveguides are special resonators with open end faces, so that not only standing waves but also travelling waves in the direction to the open end faces are possible. These waves are, however, spatially restricted in the perpendicular directions. Wave guides gain increasing importance not only in microwave technology but also in optics as optical fibers (see Sect. 12.8) and in integrated electronic circuits. We will now explore the influence of the boundary conditions for waveguides on the solutions of the wave Eq. (7.3).

7.9.1 Waves Between Two Plane Parallel Conductors

As a simple example we consider two parallel conducting planes with a distance $\Delta x = a$. Electro-magnetic waves can travel to and fro between the plates (Fig. 7.24). A wave with the electric vector $\boldsymbol{E} = \{0, \boldsymbol{E}_y, 0\}$ with the wave vector $\boldsymbol{k} = \{k_x, 0, k_z\}$ is alternately reflected by the upper ($x = a$) and the lower plane ($x = 0$) respectively.

At the reflection k_x changes its sign, while k_z does not. The wave suffers a phase change of π, so the amplitude $\boldsymbol{E}$ changes its sign. In the space between the plates we have a superposition of the waves with $\boldsymbol{k} = \{k_x, 0, k_z\}$ and $\boldsymbol{k} = \{-k_x, 0, k_z\}$. The resulting field is

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{E}_0[\sin(\omega t - k_x x - k_z z) \\ &\quad - \sin(\omega t + k_x x - k_z z)] \\ &= -2\boldsymbol{E}_0 \sin(k_x x) \cdot \cos(\omega t - k_z z) \\ \text{mit} \quad \boldsymbol{E}_0 &= \{0, E_{0y}, 0\}. \end{aligned} \tag{7.41}$$

The tangential component of the electric field $\boldsymbol{E}_t = \{0, E_y, E_z\}$ must be zero at the conducting planes $x = 0$ and $x = a$. This gives, similar to the discussion in the previous section the boundary condition

$$k_x = n \cdot \pi / a \quad (n = 1, 2, 3, \ldots). \tag{7.42}$$

Contrary to the component k_x the component k_z of the wave vector $\boldsymbol{k}$ has no restriction by boundary conditions.

Equation (7.41) describes a wave that is substantially different from standing waves (7.32a, 7.32b) resp. (7.35) because (7.41) represents a travelling wave in z-direction with its amplitude $-2E_0 \sin(k_x x)$ as a function of the coordinate x (Fig. 7.25).

The electric field $\boldsymbol{E}$ of the wave (7.41) is, according to (7.42), zero in the planes

$$x = \frac{\pi}{k_x} = \frac{a}{n}. \tag{7.43}$$

These planes are called *nodal planes* (Fig. 7.25).

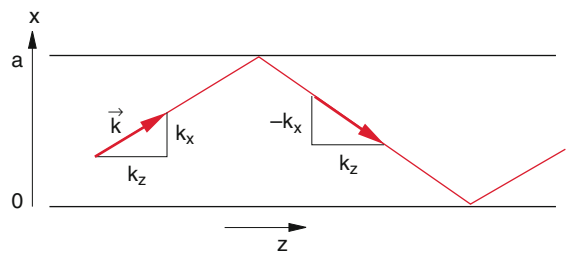


Fig. 7.24 Wave propagation wave propagation between two plane-parallel plates

Note

- Above we have discussed the special case of waves with amplitudes $\boldsymbol{E}_0 = \{0, E_{0y}, 0\}$. However with the boundary conditions $E_0 = 0$ for $x = 0$ and $x = a$ the wave Eq. (7.3) has an infinite number of further solutions with amplitudes $\boldsymbol{E}_0 = \{E_{0x}, E_{0y}, E_{0z}\}$.

Examples are:
$\boldsymbol{E} = (\boldsymbol{A} \sin k_x x + \boldsymbol{B} \cos k_x x) \cos(\omega t - k_z z)$ with $\boldsymbol{A}, \boldsymbol{B} \| \boldsymbol{e}_x$ or waves with an amplitude $\boldsymbol{E}_0 = \{0, 0, E_{0z}\}$ in z-direction.
We distinguish two types of solutions: If the electric vector is normal to the direction of propagation, i.e. $\boldsymbol{E}_0 = \{E_{0x}, E_{0y}, 0\}$ we call these waves TE-waves (transverse electrical).
If the component E_z is not equal zero, then the magnetic field $\boldsymbol{B} = \{B_{0x}, B_{0y}, 0\}$ must be orthogonal to the direction of propagation and we name the waves TM-waves (transverse magnetical) (see Sect. 7.9.2).

- The (restriction of the waves by the walls at $x = 0$ and $x = a$ results in a spatial modulation of the field amplitude in x-direction, whereas for a plane wave in z-direction in free space which is not restricted in the x or y-direction, the field amplitude is independent of x or y.

The second factor in (7.41) describes the wave $\cos(\omega t - k_z z)$ that propagates in z-direction with the phase velocity

$$v_{\text{ph}} = \frac{\omega}{k_z} \tag{7.44a}$$

Since the speed of light in vacuum is

$$c = \omega / k = \omega / \sqrt{\left(k_x^2 + k_y^2\right)^{1/2}}$$

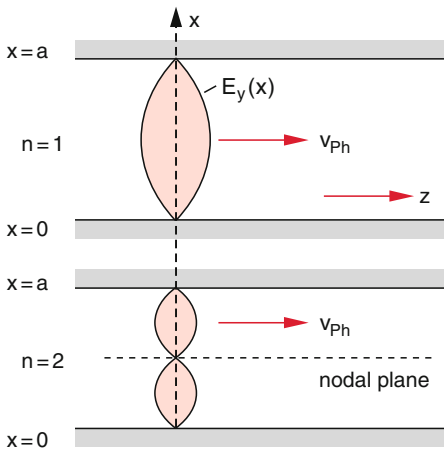


Fig. 7.25 The boundary conditions $k_x = n \cdot \pi / a$ for the k_x-component determine the electric field distribution of a wave propagating between two conductive plates at $x = 0$ and $x = a$

Equation (7.44a) can be rewritten as

$$v_{\mathrm{ph}} = \frac{c}{k_z}\sqrt{k_x^2 + k_z^2} = c \cdot \sqrt{1 + (k_x/k_z)^2} \geq c! \tag{7.44b}$$

This shows surprisingly that the phase velocity of light in a wave guide can be higher than in free space, ($v_{\mathrm{Ph}} > c$). The group velocity in wave guides

$$v_{\mathrm{G}} = \frac{\mathrm{d}\omega}{\mathrm{d}k_z} = \frac{\mathrm{d}\omega}{\mathrm{d}k} \cdot \frac{\mathrm{d}k}{\mathrm{d}k_z} = \frac{c^2}{\omega} k_z = \frac{c^2}{v_{\mathrm{ph}}} < c, \tag{7.45}$$

is, however, smaller than that for waves in free space, where for vacuum is $v_{\mathrm{G}} = v_{\mathrm{ph}} = c$.

Waves in a wave guide show dispersion, i.e. the phase velocity $v_{\mathrm{Ph}} = \omega/k$ and therefore also the group velocity v_g depend on the frequency ω (Fig. 7.26).

Using the boundary condition (7.42) $k_x = n\pi/a$ the relation $k^2 = k_x^2 + k_y^2$ yields with $k = \omega/c$

$$k_z = \sqrt{\frac{\omega^2}{c^2} - \frac{n^2\pi^2}{a^2}}, \tag{7.46}$$

so that

$$v_{\mathrm{Ph}} = \frac{c}{\sqrt{1 - \frac{n^2\pi^2c^2}{a^2\omega^2}}} \tag{7.47}$$

which shows the dependence $v_{\mathrm{Ph}}(\omega)$.

Note that v_{Ph} depends on n, i.e. the phase velocity is different for different modes.

Figure 7.26b shows the dispersion $\omega(k)$ with its slope $v_{\mathrm{G}} = \mathrm{d}\omega/\mathrm{d}k_z$ representing the group velocity.

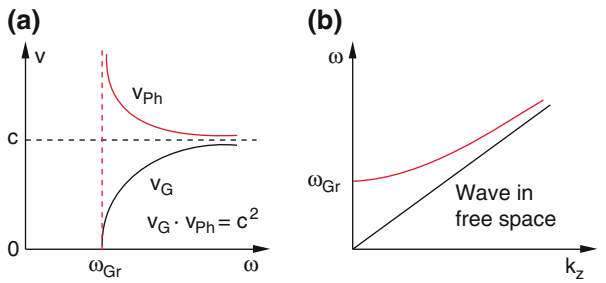


Fig. 7.26 **a**) Phase- and group velocities of electro-magnetic waves propagating between two parallel boundaries as a function of the frequency. **b**) Dispersion relation $\omega(k)$ for waves between two parallel plates (red curve) and in free space (Black cuvce). The ratio $v_{\mathrm{ph}} = \omega/k$ gives the phase velocity, while the slope $\mathrm{d}\omega/\mathrm{d}k$ gives the group velocity

Because a physical wave has a real valued component k_z of the wave vector it follows from (7.46) for the frequency

$$\omega \geq \omega_{\mathrm{G}} = n \cdot \frac{c\pi}{a} \Rightarrow v \geq v_{\mathrm{G}} = \frac{n \cdot c}{2a}. \tag{7.48}$$

We can assign to this frequency limit v_{G} an upper limit for the wavelength λ

$$\lambda \leq \lambda_{\mathrm{G}} = \frac{c}{v_{\mathrm{G}}} = \frac{2a}{n} \quad (n = 1, 2, 3, \ldots) \tag{7.49}$$

of a wave outside the wave guide.

For waves in a waveguide there exist a lower frequency limit ω_{G} and an upper limiting wavelength λ_{G}.

The wavelength λ of waves between the parallel plates cannot be larger than twice the distance a between the plates. The maximum wavelength $\lambda = 2a$ corresponds to the limiting case $k_z = 0$ for $\lambda = \lambda_{\mathrm{G}}$.

Such a wave guide acts like a filter that allows only waves with a wavelength $\lambda < \lambda_{\mathrm{G}}$ to pass. By choosing the suitable distance a we can define λ_{G} and the minimum frequency ω_{G}.

Note For $k_x = 0 \rightarrow k = k_z = \omega/c$ and $v_{\mathrm{Ph}} = c$ which implies that there is no dispersion.

Dispersion arises from the zigzag way of the wave front, due to the reflection at the walls $x = 0$ and $x = a$. The angle of $\boldsymbol{k}$ against the z-direction depends on ω.

7.9.2 Wave Guides with Rectangular Cross Section

Now we limit the space between systems of parallel plates discussed above by a further wall to realize a wave guide with rectangular cross section $\Delta x \cdot \Delta y = a \cdot b$ (Fig. 7.27), that is open in z-direction. This causes a further boundary condition in y-direction and the field amplitude $\boldsymbol{E}_0(x, y)$ of the travelling wave

$$\boldsymbol{E}(x, y, z, t) = \boldsymbol{E}_0(x, y) \cdot \cos(\omega t - k_z z) \tag{7.50}$$

becomes a function of x and y.

At the conducting walls the tangential component of $\boldsymbol{E}$ must be zero.

Using the ansatz (7.50) in the wave Eq. (7.3a) yields

$$\frac{\partial^2 \boldsymbol{E}}{\partial x^2} + \frac{\partial^2 \boldsymbol{E}}{\partial y^2} + \left(\frac{\omega^2}{c^2} - k_z^2\right) \cdot \boldsymbol{E} = 0. \tag{7.51}$$

In analogy to the wave guides that are restricted only in x-direction, we now get two types of solutions: transverse electric TE-waves with $\boldsymbol{E} = \{E_x, E_y, 0\}$ and transverse magnetic TM-waves with $\boldsymbol{B} = \{B_x, B_y, 0\}$.

The general solution (7.50) can be found from (7.51) by using Maxwell's equations. With the boundary conditions

$$k_x = n\pi/a; \quad k_y = m\pi/b \tag{7.51a}$$

these are the solutions for TE-waves

$$\begin{aligned} E_{0x}(x,y) &= A \cdot \cos\frac{n\pi}{a}x \cdot \sin\frac{m\pi}{b}y, \\ E_{0y}(x,y) &= B \cdot \sin\frac{n\pi}{a}x \cdot \cos\frac{m\pi}{b}y, \\ E_{0z} &= 0. \end{aligned} \tag{7.52}$$

The magnetic field is then obtained from

$$\mathbf{rot}\, \boldsymbol{E} = -\frac{\partial \boldsymbol{B}}{\partial t}. \tag{7.52a}$$

We consider as example, a special TE_{nm}-solution with $E_x = E_z = 0$ and $n = 1, m = 0$ illustrated in Fig. 7.27. From (7.52) and (7.50) it follows with $k_x = \pi/a$:

$$E_y = E_0 \sin\left(\frac{\pi}{a}x\right)\cos(\omega t - k_z z). \tag{7.50a}$$

The magnetic field of this so called TE_{10}-wave is obtained with the help of (7.52a):

$$\begin{aligned} B_x &= -\frac{k_z}{\omega}E_0 \sin(k_x x)\cdot\cos(\omega t - k_z z), \\ B_y &= 0, \\ B_z &= -\frac{k_x}{\omega}E_0 \cos(k_x x)\cdot\sin(\omega t - k_z z). \end{aligned} \tag{7.53}$$

One observes that $\boldsymbol{B} = \{B_x, 0, B_z\}$ is no longer normal to the direction of propagation z, (Fig. 7.27x), because the magnetic field has a component $B_z \neq 0$.

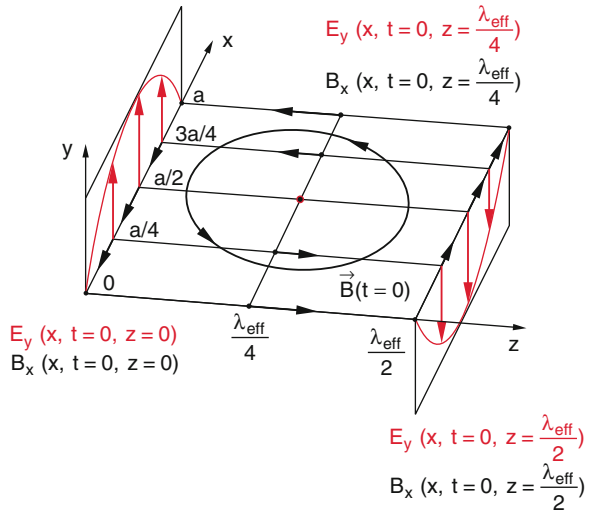


Fig. 7.27 Waveguide with rectangular cross section and with a $TE_{1,0}$ wave ($\boldsymbol{E} = \{0, E_y, 0\}$ and $\boldsymbol{B} = \{B_x, 0, B_z\}$) propagating into the z-direction

Figure 7.27 shows a snapshot of the electric and magnetic fields at time $t = 0$ in the three planes $z_0 = 0, z_1 = \frac{1}{4}\lambda_{\text{eff}}, z_2 = \frac{1}{2}\lambda_{\text{eff}}$.

For the electric field we obtain

$$\begin{aligned} E_y(x, z_0) &= E_0 \sin\frac{\pi}{a}x \\ E_y(x, z_1) &= 0 \\ E_y(x, z_2) &= -E_0 \sin\frac{\pi}{a}x \end{aligned}$$

and for the magnetic field in the plane $z_0 = 0$

$$\begin{aligned} B_x(x = 0, z_0) &= 0 \\ B_x(x = a, z_0) &= 0 \\ B_x(x = a/4, z_0) &= B_x(x = 3a/4, z_0) \\ &= -\frac{1}{2}\sqrt{2}k_z\frac{E_0}{\omega} \\ B_x(x = a/2, z_0) &= -k_z\frac{E_0}{\omega} \\ B_z(x, z_0) &= 0, \end{aligned}$$

Whereas for the magnetic field in the plane $z_1 = \lambda/4$ we get

$$\begin{aligned} B_x(x, z_1) &= 0 \\ B_z(x = 0, z_1) &= -B_z(x = a, z_1) \\ &= \pi\frac{E_0}{\omega a} \\ B_x(x = a/4, z_1) &= -B_z(x = 3a/4, z_1) \\ &= \frac{1}{2}\sqrt{2}\pi\frac{E_0}{\omega a} \\ B_z(x = a/2, z_1) &= 0. \end{aligned}$$

The values for the field at plane z_2 correspond to those at z_0 but with the opposite sign.

For the wave vector in the direction of propagation, we get by inserting (7.50) into (7.51) the condition

$$k_x^2 E_y + k_z^2 E_y - \frac{\omega^2}{c^2}E_y = 0,$$

from which follows

$$k_z = \sqrt{(\omega^2/c^2) - \pi^2/a^2}. \tag{7.54a}$$

The effective wavelength $\lambda_{\text{eff}} = 2\pi v_{\text{Ph}}/\omega$ with $v_{\text{Ph}} = \omega/k_z$ results from (7.47) as:

$$\lambda_{\text{eff}} = \frac{\lambda_0}{\sqrt{1 - (\lambda_0/2a)^2}}, \tag{7.54b}$$

if $\lambda_0 = c/\nu$ is the wavelength of a wave with the same frequency but in free space.

The wavelength of a cavity wave is larger than in free space!

In wave guides not only TE-waves exist, but also TM-waves with a transverse magnetic field while the electric field has a component in the direction of propagation (Fig. 7.28).

The corresponding solutions of the wave equation are e.g.

$$\begin{aligned} E_x &= E_0 \frac{k_x k_z}{k_x^2 + k_y^2} \cos(k_x x) \sin(k_y y) \\ &\quad \cdot \sin(\omega t - k_z z); \\ E_y &= E_0 \frac{k_y k_z}{k_x^2 + k_y^2} \sin(k_x x) \cos(k_y y) \\ &\quad \cdot \sin(\omega t - k_z z); \\ E_z &= E_0 \sin(k_x x) \sin(k_y y) \cos(\omega t - k_z z); \end{aligned} \tag{7.55a}$$

$$\begin{aligned} B_x &= -\frac{\omega}{k_z c^2} E_y \\ B_y &= +\frac{\omega}{k_z c^2} E_x \\ B_z &= 0. \end{aligned} \tag{7.55b}$$

with the corresponding boundary condition (7.51a) for k_x and k_y, [13].

Such travelling waves in wave guides are called TE_{nm}- resp. TM_{nm}-waves, depending on whether $\boldsymbol{E}$ or $\boldsymbol{B}$ is orthogonal to the z-direction. With the boundary condition (7.51a) the spatial distribution of the amplitudes in the xy-plane has n nodal x-planes at $x = x_n$ and m nodal y-planes at $y = y_m$.

For TE-waves it is possible, that n or m are zero. For TM-waves n as well as m must be greater than zero. So the TM_n-wave, shown in Fig. 7.27, is the most simple TM-wave.

In semiconductors waveguides with circular cross section, we derive from its boundary conditions for the pre factors Bessel functions instead of $\sin(k_x x)$ and $\cos(k_y y)$.

A few examples for such distributions of the amplitudes $E(x, y)$ in cylindrical cavities are shown in Fig. 7.29. In this case is n the number of radial and m the number of the azimuthal nodal surfaces.

The limiting frequency ω is determined analogous to (7.48). From the relation

$$k_z = \sqrt{k^2 - k_x^2 - k_y^2}$$

We obtain with $k = \omega/c$ and the boundary conditions (7.42) the relation

$$\omega \geq \omega_G = c \cdot \pi \sqrt{\frac{n^2}{a^2} + \frac{m^2}{b^2}}. \tag{7.56}$$

If we choose suitable values for a and b we can, for example, achieve that only one TE_n-wave mode is realized for a fixed frequency ω.

Such wave guides play an important role for the transmission of microwaves [12]. They prohibit that microwave power produced by a transmitter, radiates into all directions.

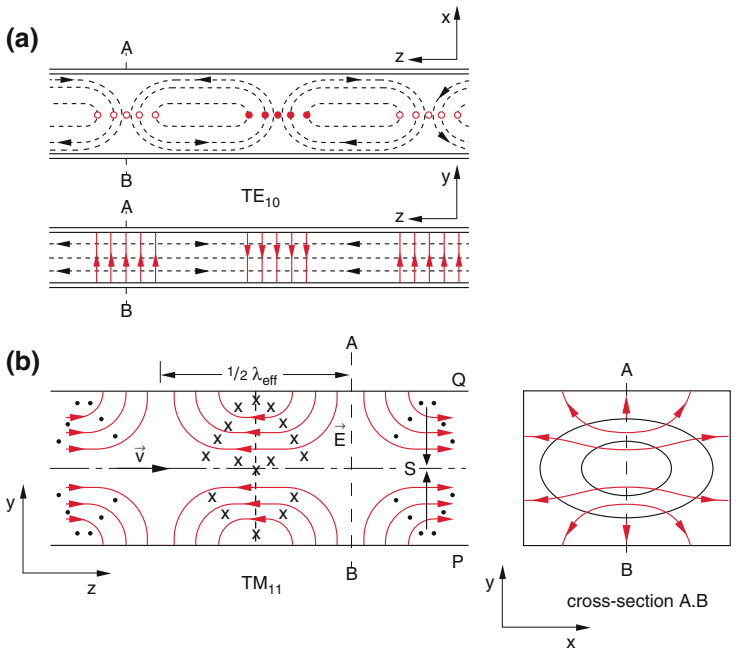


Fig. 7.28 TEM-waves in a waveguide with rectangular cross section. The red lines are the electric field lines **a**) TE_{10} wave. The electric field lines point into the $\pm y$-direction, **b**) TM_{11}-wave The magnetic field is perpendicular to the drawing plane. ($\times$ means $\boldsymbol{B}$ points into the drawing plane, • means $\boldsymbol{B}$ points out of the drawing plane)

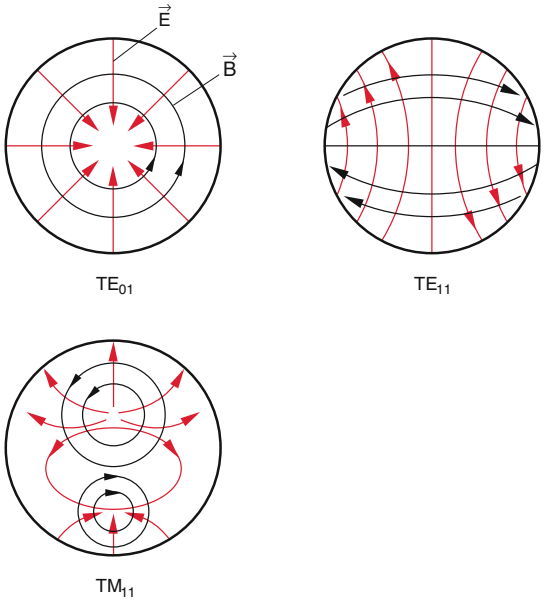


Fig. 7.29 Examples for the field distribution of TM_{nm} and TE_{nm} in a cylindrical waveguide with circular cross section

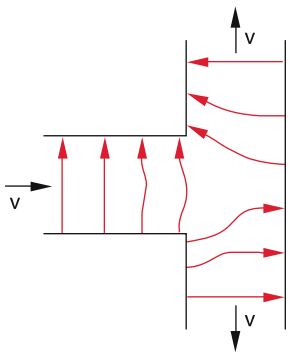


Fig. 7.30 Branching of a waveguide

The conducting walls create a finite volume where the waves are "guided", practically without loss, to the location of consumption. With this arrangement we can guide waves "around the corner" (Fig. 7.30). This possibility allows a great variability of wave guides.

The commercially available cavity resonators consist of conducting parts with flanges (Fig. 7.31) so that by combining various parts the wave guides can be adapted to the problem at hand.

Fig. 7.31 Commercial microwave guides with T-part and flanges

7.9.3 Waves Along Wires; Lecher Line; Coaxial Cable

Electromagnetic waves can not only be "guided" by cavity resonators, but can be also transmitted along electrical conducting wires and in coaxial cables. We illustrate this with the equipment shown in Fig. 7.32.

7.9.3.1 Lecher-Line

Two parallel wires in z-direction connected with each other on one end (*Lecher line*) are placed in an electromagnetic field produced by a high frequency transmitter.

Now we observe standing waves along the wires that create a spatially periodic voltage $U(z)$ and a corresponding current distribution $I(z)$.

The distribution of the voltage $U(z)$ can be demonstrated by a glow lamp or a sensitive light bulb that is connected to the two wires and can slide along the z-axis (Fig. 7.32a).

At the open end of the Lecher line is a maximum of the voltage and a node at the short circuited end.

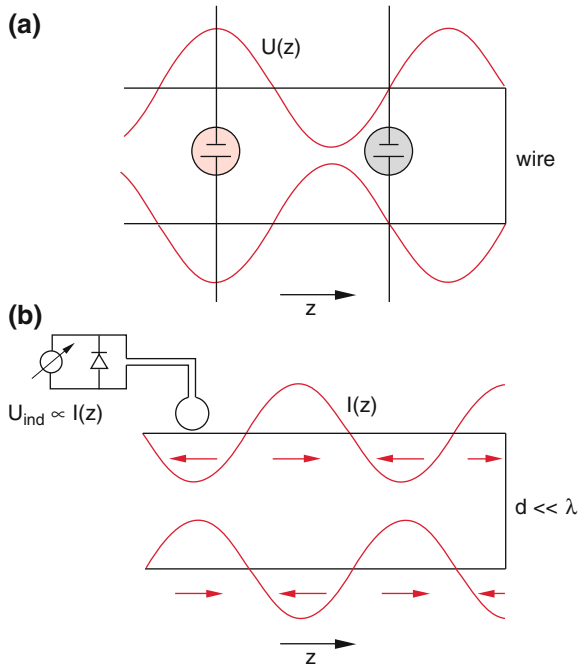


Fig. 7.32 Lecher conductors **a**) measurement of the voltage $U(z)$ with a glow lamp dipole antenna, **b**) measurement of the current $I(z)$ with an induction coil. The current has amplitude maxima at the shortened end and nodes at the open end, for the voltage $U(z)$ the situation is just reversed. The distance d is here enlarged in order to make the illustration more clear

The distribution of the current $I(z)$ can be shown by its magnetic field $B(r, z)$ that induces a voltage in a small coil above the conductors With a rectifying diode the induced voltage can be measured directly by a voltmeter (Fig. 7.32b).

The current $I(z)$ is zero at the open end of the Lecher line and has a maximum at the closed end, where a phase shift of π occurs. If the distance between the two conductors becomes small compared to the wavelength λ of the standing wave, then the currents in the both conductors have opposite phases.

7.9.3.2 Coaxial Cables

In Chap. 6 we have learned that a straight wire that carries an alternating current of high frequency acts like a Hertzian dipole and radiates energy in the form of electromagnetic waves. The transmitted power is according to (6.38) proportional to the fourth power of the frequency ω.

Therefore the transport of high frequency electric currents through single wires is not practicable, because of the large energy losses. In this case twin-conductors can be used (Fig. 7.32), where the distance between both conductors is small compared to the wavelength λ, because then the waves radiated from the two wires have a phase shift of π and therefore interfere destructively.

Still better for the reduction of the losses by radiation is a coaxial cable (Fig. 1.21). It consists of a thin wire as inner conductor with radius a and a concentric outer conductor with radius b (Fig. 7.33). This system can be regard as cylindrical wave guide with circular cross section.

The essential difference to the normal wave guides is the inner conductor that represents an additional boundary condition. If the outer conductor is grounded, the electric field is radial. Direction and amount of the electric field $\boldsymbol{E}$ depend on the potential V of the inner conductor. The magnetic field lines are concentric circles about the inner conductor The direction of rotation of the magnetic field as a function of z changes periodically from clockwise to counterclockwise with the wavelength λ as period.

If an electromagnetic wave is travelling in z-direction through the coaxial cable, the voltage U between inner and outer conductor becomes a function of z (Fig. 7.34).

With the inductance $\hat{L}$ and the capacitance C per unit length of the cable, we get for the voltage change along the cable, according to the law of induction

$$\Delta U = U(z+\Delta z) - U(z) = -\hat{L}\Delta z \frac{\mathrm{d}I}{\mathrm{d}t},$$

Which yields for $\Delta z \to 0$

$$\frac{\partial U}{\partial z} = -\hat{L}\frac{\partial I}{\partial t}. \tag{7.57}$$

The electric charge per cable length Δz is

$$Q = \hat{C} \cdot U \cdot \Delta z.$$

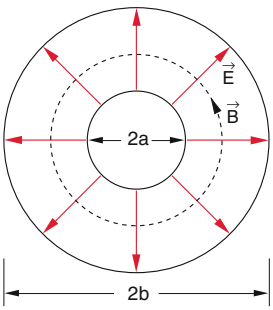


Fig. 7.33 Coaxial waveguide with radial electric field lines and circles of the magnetic field lines

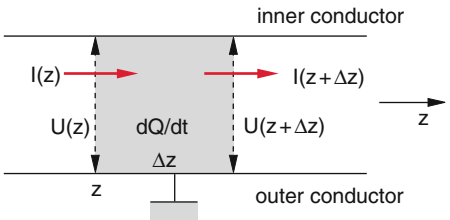


Fig. 7.34 Illustration of the derivation of the wave Eqs. (7.59a, 7.59b)

The temporal change $\partial Q/\partial t$ corresponds to the current

$$\Delta I = I(z+\Delta z) - I(z),$$

flowing out of or into the volume between z and $z + \Delta z$. Therefore it is

$$\frac{\partial I}{\partial z} = -\hat{C}\frac{\partial U}{\partial t}. \tag{7.58}$$

Differentiating (7.57) with respect to z and (7.58) with respect to t and inserting $\partial^2 I/(\partial z \partial t)$ into (7.58) yields the equations

$$\frac{\partial^2 U}{\partial z^2} = \hat{L}\hat{C}\frac{\partial^2 U}{\partial t^2}, \tag{7.59a}$$

$$\frac{\partial^2 I}{\partial z^2} = \hat{L}\hat{C}\frac{\partial^2 I}{\partial t^2}. \tag{7.59b}$$

These are wave equations for the voltage $U = U_0 \cdot \sin(\omega t - kz)$ and the current $I = I_0 \cdot \sin(\omega t - kz - \varphi)$. The amplitudes of voltage and current propagate into the z-direction with the speed

$$v_{\mathrm{Ph}} = \frac{1}{\sqrt{\hat{L}\cdot\hat{C}}} \tag{7.60}$$

In general the resistance $Z = U/I$ is a complex quantity which depends on the phase shift between U and I. We get the relation (see Sect. 5.4)

$$\tan\varphi = \frac{\mathrm{Im}(Z)}{\mathrm{Re}(Z)}.$$

The real quotient $Z_0 = U_0/I_0 = \sqrt{\hat{L}/\hat{C}}$ is called **wave resistance** of the coaxial cable (see Problem 7.15). Its unit is $[Z_0]$ = 1 V/A= 1 Ω.

If we connect a resistor $R = Z_0$ to one end of the cable then the wave travelling through the coaxial cable will not be reflected.

Example

A coaxial cable with $\hat{C} = 100\,\mathrm{pF/m}$ and $\hat{L} = 0.25\,\mu\mathrm{H/m}$ has a wave resistance of Z_0 = 50 Ω.

The wave resistance of a coaxial cable as that shown in Fig. 7.33 depends on the radii a and b of the inner and outer conductor. We get (see Problem 7.15)

$$Z_0 = \frac{1}{2\pi\varepsilon_0 c}\ln(b/a). \tag{7.61}$$

In a flexible coaxial cable is the inner conductor a thin wire, the outer conductor a wire net. The space between inner and outer conductor is filled with an insulator ($\varepsilon > 1$). This increases the capacitance C of the cable by the factor ε and the phase velocity is therefore smaller by the factor ε than in vacuum (see Chap. 8).

In coaxial cables as well as in free space the fields E and B are transverse to the direction of propagation. The waveforms of the waves travelling in the cable are called TEM_{nm}-modes (transverse electromagnetic). They have n nodes in r-direction and m nodes along the azimutal coordinate.

7.9.4 Examples of Wave Guides

Now we will consider some examples of wave guides for different wavelengths.

7.9.4.1 Radiowaves in the Atmosphere of the Earth

The radiation of the sun with very short wavelengths ionizes part of the molecules and atoms in the higher earth atmosphere in altitudes above 50–100 km.

This ionosphere reflects electromagnetic waves in the rf-range (radio frequency). In the layer where the dielectric constant ε changes rapidly, reflection of the wave takes place. This layer is named *Heaviside layer* (Fig. 7.35). The frequency limit ν_G depends on the ionization density which in turn depends on the season of the year, on day and night and on the activity of the sun.

The layers are divided in the D-layer (approx. 80 km), E-layer (120 km) and F-layer (200–400 km).

The D-layer reflects waves with frequencies $\nu > 5$–30 MHz. Because of this reflections, waves from the transmitter S can reach points B the earth, which cannot be reached by mere geometric propagation.

The ionosphere and such the height of the Heaviside layer are influenced by the flux of particles from the sun and therefore changes of this flux by protuberances or flares on the surface of the sun can change the reflectivity of the ionosphere and disturb the radio communication.

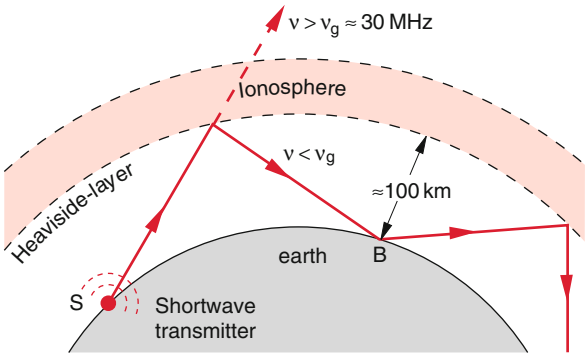


Fig. 7.35 Reflection of radio waves at the Heaviside layer between ionosphere and mesosphere

7.9.4.2 Microwave Guides

In Sect. 7.9.2 we have discussed metallic cavities (conductors). In microwave spectroscopy they are used to transport microwaves over short distances from the transmitter through the absorbing sample to the detector. This is a common method to guide microwaves over distances of several meters.

7.9.4.3 Wave Guides for Light

Also light waves can be transported for distances up to 1000 km by thin, flexible fibers of quartz. Today these optical fibers are used to transmit digital optical signals with bit rates up to 10^{11} s^{-1} (see Sect. 12.7).

7.10 The Electromagnetic Frequency Spectrum

The Maxwell equations and the wave equation derived from them describe electromagnetic fields and their transmission as waves. Periodic waves are special cases of many other possible solutions of the wave equation. The frequency ω and the wavelength $\lambda = 2\pi c/\omega$ of these solutions are still undefined and can be selected arbitrarily.

All phenomena of electromagnetic waves in vacuum like for instance, the speed of propagation c, the energy density w_{em}, The Poynting vector $\boldsymbol{S}$, must be described by the Maxwell equations for all frequencies.

The whole frequency range of electromagnetic waves known to us today comprises a range of 24 decades. In order to readily compare the corresponding frequencies ν (in Hz), wavelengths λ (in m) and photon energies (in eV) a schematic comparison is illustrated in Fig. 7.36.

As shown in quantum physics, a photon with energy $h \cdot \nu$ is the smallest unit of energy of an electromagnetic field with frequency ν. The energy density w_{em} of the electromagnetic field is quantized and can be always written as $n \cdot h \cdot \nu$, where n is the number of photons per unit volume. The constant h is Planck's constant (see Vol. 3, Sect. 3.1).

The total spectral range of the electromagnetic field in vacuum can be described by the four Maxwell equations and the constants ε_0, μ_0.

The situation changes fundamentally if matter is present, because now the interaction between electromagnetic field and matter has to be considered. Then the frequency dependent properties of matter become important, i.e. absorption, scattering, dispersion or reflection (see Chap. 8).

The investigation of the material properties in different ranges of the spectrum has brought an enormous increase of our knowledge about the subatomic structure of matter (see Vols. 3, and 4).

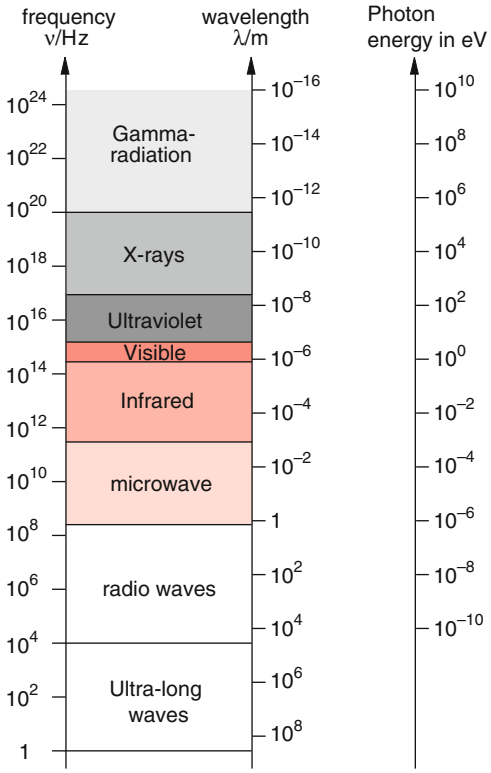


Fig. 7.36 Survey about the whole electromagnetic spectrum known up to now

Up to about 1910 investigations were only possible in the visible range of electromagnetic waves (λ = 400–700 nm), because the human eye was the only known detector of electromagnetic waves. The phenomena and their description, found in that range of visible light are the subject of optics. In the meantime a number of radiation sources and detectors have been developed that work in much more extended ranges beyond the visible spectrum. These are:

The *infrared spectral range* 700 nm < λ < 100 μm
The microwave range 100 μm < λ < several cm
the visible range 400 nm < λ < 700 nm
The ultraviolet range 10 nm < λ < 400 nm
The X-ray range 0.01 nm < λ < 10 nm
The gamma radiation range λ < 0.01 nm.

For all these ranges new experimental methods have been developed. Therefore modern optics includes as well the infrared and ultraviolet range.

Especially remarkable are the advances in astrophysics by disclosing new areas of the spectrum. In former times the

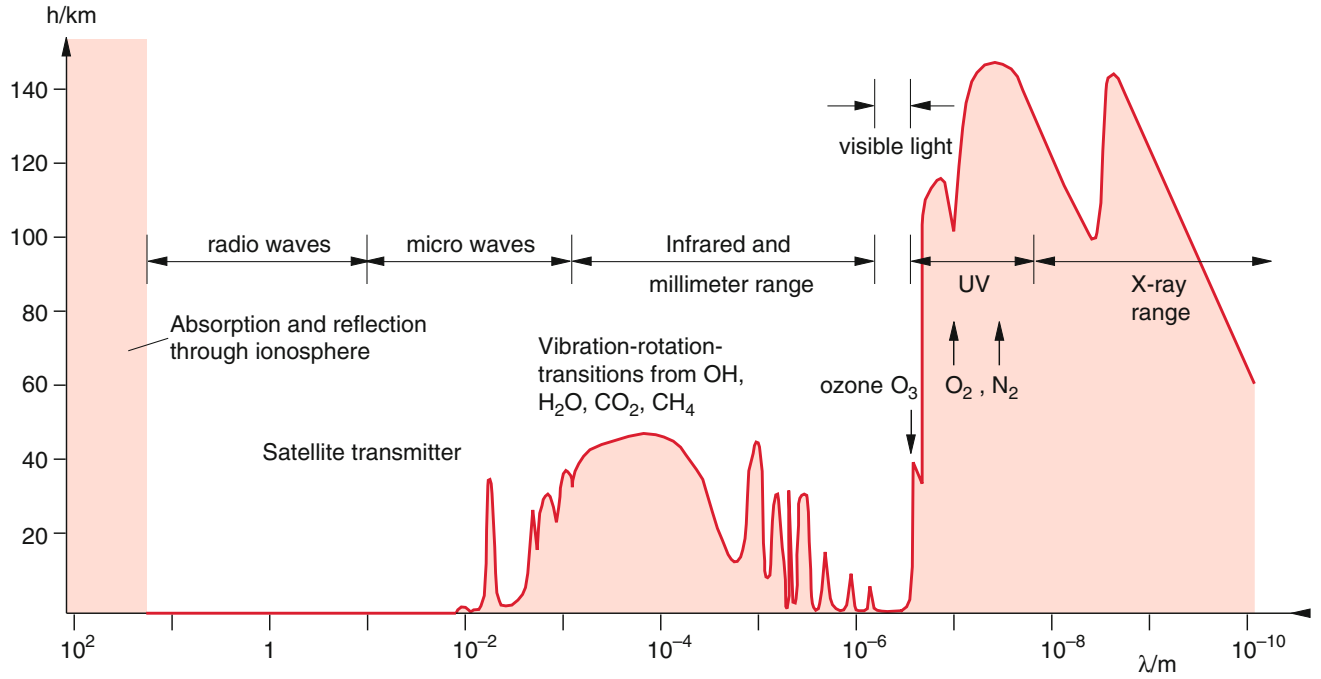


Fig. 7.37 Spectral absorption of the earth atmosphere. The red cure gives the height h above the earth surface where the intensity of the radiation, incident from outside, has decreased to $1/e$ of its initial value. This illustrates that there are only a few spectral windows $\Delta\lambda$ where the intensity $I(\lambda)$ reaches the ground without significant attenuation

observation of celestial bodies (planets, comets, stars or galaxies) by astronomers was restricted to visible light, except for the investigation of meteors. In the meantime radio astronomy has brought a wealth of new information. Also in the infrared and X-ray range, previously unknown phenomena are now accessible. Besides observations from satellites outside of the earth atmosphere also observers bound to the surface of the earth can watch the sky using electromagnetic waves with frequencies that are not absorbed by the atmosphere. These are mainly visible light, radio frequencies and small windows in the near infrared, where the atmosphere is less absorbing, so that radiation can reach the surface of the earth. In Fig. 7.37 the spectral windows accessible to observers on the earth are shown. The red curve gives that altitude above the earth surface where the radiation from outside has dropped due to absorption to $1/e$ of its value outside the atmosphere. The radiation with wavelengths $\lambda > 180$ nm is mainly absorbed by trace gases in the atmosphere such as CO_2, H_2O or CH_4.

The main components of the atmosphere, N_2 and O_2, absorb only in the spectral range with $\lambda < 180$ nm, in the so called vacuum-ultraviolet range. Therefore measurements in this range have to be done outside the atmosphere, i.e. from balloons, satellites or stations in the orbit.

The transmitters for the various ranges of the electromagnetic spectrum differ very much. For waves with $\lambda > 1$ m ($\nu < 300$ MHz) radio frequency transmitters are commercially available (see Fig. 6.17). For microwaves $\lambda > 1$ mm, $\nu < 300$ GHz), sources e.g. clystrons or carcinotrons [13, 14] are also commercially available. Infrared radiation is emitted by thermal sources, e.g. wolfram filaments heated to about 2000 K. For visible light is the transition between energy levels of atoms and molecules the main source of radiation. (see Vol. 3).

Because of the importance for the human being, the next chapters deal with phenomena and interactions of visible light with matter, superposition of electromagnetic waves (interference or diffraction). The phenomena are demonstrated for light, because these waves are essentially conspicuous and can be observed without an additional detector.

The principles and phenomena of visible light discussed in the following chapters, are the subject of optics. Today it includes also the infrared and ultraviolet range of the spectrum, because many detectors for these ranges are nowadays available.

Summary

- All electromagnetic waves in vacuum are solutions of the wave Eq. (7.3)

$$\Delta \boldsymbol{E} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2}.$$

  which can be derived from the Maxwell equations.
- Plane periodic waves $\boldsymbol{E} = \boldsymbol{E}_0 \cos(\omega t - \boldsymbol{k} \cdot \boldsymbol{r})$ are important special cases of the general solutions.
- The speed of propagation $c = \omega k$ of electromagnetic waves in vacuum is the same for all frequencies ω, i.e. there is no dispersion. The value of $c = 299{,}792{,}458$ m/s is now defined and is used for the definition of the unit of length. 1 m (m) is that distance which is traveled by light in the time $t = 1/299{,}792{,}458$ s.
- Between electric and magnetic field of an electromagnetic wavet the relations hold:

$$|\boldsymbol{E}| = c|\boldsymbol{B}|; \quad \boldsymbol{E} \perp \boldsymbol{B}; \quad \boldsymbol{E}, \boldsymbol{B} \perp \boldsymbol{k}.$$

  where $\boldsymbol{E}$, $\boldsymbol{B}$ and $\boldsymbol{k}$ form a right-handed system.
- The electromagnetic wave transports energy and momentum. The Poynting vector

$$\boldsymbol{S} = \varepsilon_0 c^2 (\boldsymbol{E} \times \boldsymbol{B})$$

  gives the direction of transport.
- The intensity I of a wave is the energy transported per unit time (second) through the unit area (square meter). The following relations pertain:

$$I = |\boldsymbol{S}|.$$

The momentum of the electromagnetic wave per unit volume is

$$\pi_{\mathrm{St}} = \frac{1}{c^2}\boldsymbol{S}.$$

- The stationary solutions of the wave equation in closed resonators are standing waves $\boldsymbol{E} = \boldsymbol{E}_0 \sin(\boldsymbol{k} \cdot \boldsymbol{r})$. The spatial distribution of their amplitudes is determined by the boundary conditions at the walls of the resonator.
- In wave guides with propagation possibilities in z-direction TE_{mn}-respectively TM_{mn}-waves $\boldsymbol{E} = \boldsymbol{E}_0(x, y) \cdot \cos(\omega t - k_z z)$, with amplitudes depending on x and y, propagate into the z-direction.

Problems

7.1 Prove, that from the Maxwell Eqs. (7.1a, 7.1b) a wave equation for the magnetic field B can be derived, that is analogue to the Eq. (7.3) for the electric field.

7.2 Show, that for an arbitrary plane wave, propagating into the direction of $\boldsymbol{k}$, the planes $\boldsymbol{k} \cdot \boldsymbol{r}$ = constant are phase planes.

7.3 From the linearity of the wave equation it follows that every linear combination of solutions for the field amplitudes represents the most general solution. Is this also valid for the intensities? Are there situations where the sum of the intensities of two waves gives the total intensity?

7.4 Prove that every linear polarized wave can be composed of two circular polarized waves with opposite directions of rotation.

7.5 A sun energy collector has an effective area of 4 m^2 and consists of a blackened metal plate which absorbs 80% of the incident energy. The plate has in its interior pipes for the cooling water. What is the maximum water flow, if the temperature of the plate should not exceed 80 °C? The unwanted energy exchange with the surroundings (T = 20 °C) is $\Delta Q = \kappa \cdot \Delta T$ with κ = 2 W/K. The incident sun radiation is 500 W/m^2 and hits the collector under an angle of 20° against the surface normal.

7.6 A capacitor with plane parallel plates has the capacity C. it is charged by the constant current $I = dQ/dt$.

(a) Determine the electric and the magnetic field during the charging.

(b) How large is the Poynting vector $\boldsymbol{S}$

(c) Express the total energy that is used to transport the charge $\boldsymbol{Q}$ to the capacitor by the Poynting vector and by the charge $\boldsymbol{Q}$ and capacitance $\boldsymbol{C}$.

7.7 The sun radiates to the earth (outside the atmosphere) the intensity I = 1400 W/m^2. How much sun energy receives Mars?

Assume Mars diffusely reflects 50% of the incident energy (i. e uniformly into the angle 2π). How much of this energy can the earth receive at the time when the earth is located between sun and Mars

(distance earth-sun = 1.5 × 10^{11} m and sun-Mars = 2.3 × 10^{11} m).

7.8 The maximum intensity of the solar radiation falling onto an area perpendicular to the earth surface is in Germany in June about 800 W/m^2. Which radiation power would pass through the pupil of the eye with a diameter of 2 mm, if one looks without filters into the sun? The eye lens images the sun onto the retina and forms there a spot with 0.1 mm diameter. What is the intensity on the retina?

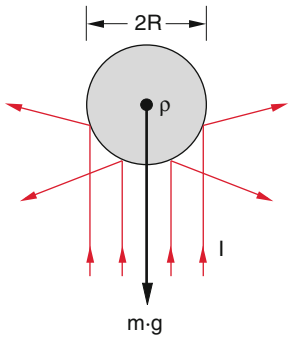


Fig. 7.38 To problem 7.9

7.9 A small ball can float in air, if the radiation pressure caused by a vertical laser beam just balances the gravitation (Fig. 7.38). How large must be the intensity of the laser beam with a constant intensity over the cross section of the beam which is equal to that of the ball. The reflectivity of the ball is 100%.

7.10 (a) A light mill in vacuum with 4 quadratic wings (2 × 2 cm^2) consisting alternately of totally absorbing resp. reflecting surfaces is hit by a light beam with a cross section of 6 × 6 cm^2 and an intensity of 10^4 W/m^2. How large is the torque acting on the light mill?

(b) Now the mill is brought into a container filled with argon (p = 10 mbar). The heat capacity of the absorbing surfaces is 10^{-1} Ws/K. Estimate the torque for a gas temperature of 20 °C.

7.11 A small antenna emits a radiation power of 1 W, which is collected by a parabolic mirror with 1 m diameter and a focal length of 0.5 m, which reflects the radiation as a parallel beam (plane wave). What is the intensity of the plane wave, if the dipole sits in the focal point of the parabolic mirror and its axis is orientated perpendicular to the line SO in Fig. 7.39?

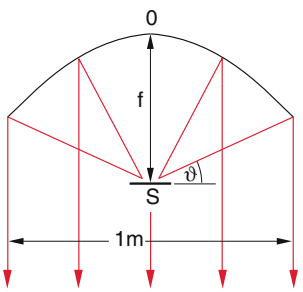


Fig. 7.39 To problem 7.11

7.12 In a wave guide with quadratic cross section (a = 3 cm) an electromagnetic wave propagates with the group velocity $v_G = 10^8$ m/s. What is the value of its wavelength λ and how large is the phase velocity v_{Ph}?

7.13 A current of 30 A flows through a straight copper wire (3 mm diameter, resistance R = 0.03 Ω/m length L = 100 m). Calculate $\boldsymbol{E}$, $\boldsymbol{B}$, and the Pointing vector $\boldsymbol{S}$ at the surface of the wire.

7.14 There are plans to send space ships on the journey to other planets. They shall be accelerated by photon recoil to large velocities. How large must be the intensity of a light source with 100 cm^2 area in the space ship, which sends light backwards to produce the recoil, in order to reach an acceleration of 10^{-5} m/s^2 for a mass of 1000 kg?

7.15 Calculate for a coaxial wave guide with inner radius a and outer radius b the capacitance per m and the wave resistance Z_0. How large must be b for a = 1 mm in order to reach a value Z_0 = 100 Ω?

References

1. https://en.wikipedia.org/wiki/Crookes_radiometer
2. Julio A. Fernandez: Comets, Nature, Dynamics Origin and their Cosmogonical Relevance (Springer Heidelberg 2005)
3. Hanss Zischler: The Comet: The Journey of Rosetta (editions Xavier Barral, 2019)
4. O. Rømer: Philos. Trans. 12, 893 (June 25,1677)
5. F. Tuinstra: Rømer and the finite speed of light. Phys. Today 57, Dec.2004 p-16–17
6. H. Fizeau H. Sur les hypothèses relatives à l'éther lumineux. In: Comptes Rendus. 33, 1851, S. 349–355
7. Sur les vitesses relatives de la lumière dans l'air et dans l'eau / par Léon Foucault. Thèse présentée 1853 à la faculté des sciences de Paris. p. 36
8. Th. Udem, R. Holzwarth, T. W. Hänsch, Optical Frequency Metrology. Nature 416, 233 (2002)
9. Hänsch, Theodor W. (2006). "Nobel Lecture: Passion for precision". *Reviews of Modern Physics*. **78** (4): 1297–1309
10. Hall, John L. (2006). "Nobel Lecture: Defining and measuring optical frequencies". *Reviews of Modern Physics*. **78** (4): 1279–1295
11. NIST, fundamental constants CODATA https://physics.nist.gov/cuu/Constants/
12. https://en.wikipedia.org/wiki/Resonator#Cavity_resonators
13. https://en.wikipedia.org/wiki/Waveguide_(electromagnetism)
14. http://www.radartutorial.eu/08.transmitters/Traveling%20Wave%20Tube.en.html

Electromagnetic Waves in Matter

8

In the previous chapter we have discussed the characteristics of electromagnetic waves in vacuum. We will now investigate the influence of matter on the propagation of electromagnetic waves. For this purpose we have to add to the Maxwell Eq. (7.1), which are valid for waves in vacuum, terms that describe the different effects of matter on electromagnetic waves.

While the propagation and superposition of electromagnetic waves in matter can be satisfactorily treated by a classical macroscopic theory, based on the Maxwell equations, the generation and annihilation of electromagnetic waves (emission and absorption) by the atoms or molecules of the medium can be only described quantitatively by a microscopic model based on quantum theory (see Vol. 3).

Nevertheless the classical model of the damped oscillator for the absorbing or emitting atoms, which we have already used for the description of the Hertzian dipole, can give a good intuitive insight into the physical phenomena observed for electromagnetic waves in matter.

We will at first give a vivid phenomenological representation of these phenomena, before we treat the solutions of the extended Maxwell equations.

8.1 Refractive Index

Measuring the phase velocity v_{ph} of electromagnetic waves in matter we find:

- The phase velocity v_{ph} in matter is smaller than that in vacuum by a factor n that depends on the medium

$$v_{ph}(n) = \frac{c}{n}. \tag{8.1}$$

- The value of n and therefore also that of v_{ph} not only depends on the medium but also on the wavelength λ of the wave.

$$n = n(\lambda) \;\rightarrow\; v_{ph} = v_{ph}(\lambda)\,(\textit{dispersion}).$$

How can we understand these results?
We regard in Fig. 8.1 a plane wave

$$\boldsymbol{E}_e = \boldsymbol{E}_0 e^{i(\omega t - kz)} = \boldsymbol{E}_0 e^{i\omega(t - z/v_{Ph})},$$

which travels in the z-direction through a medium (for instance a gas layer) with the thickness Δz. Inside the medium the wavelength $\lambda = \lambda_0/n$ is smaller than the wavelength λ_0 in vacuum, whereas the frequency ω is the same inside and outside.

The electromagnetic wave induces the atomic electrons to forced oscillations. They can be regarded as oscillating dipoles that emit again electromagnetic waves with the same frequency as the exciting wave. However, the phase of the

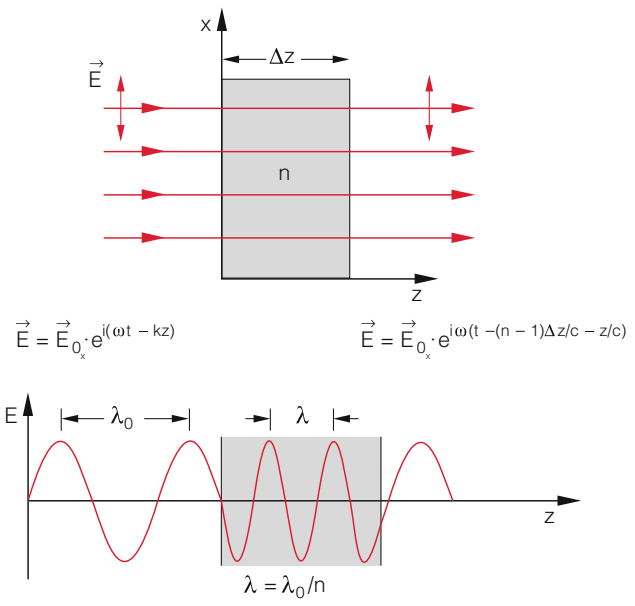


Fig. 8.1 Passage of a plane wave through a medium with refractive index n. The reflection at the interfaces has been here neglected

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_8

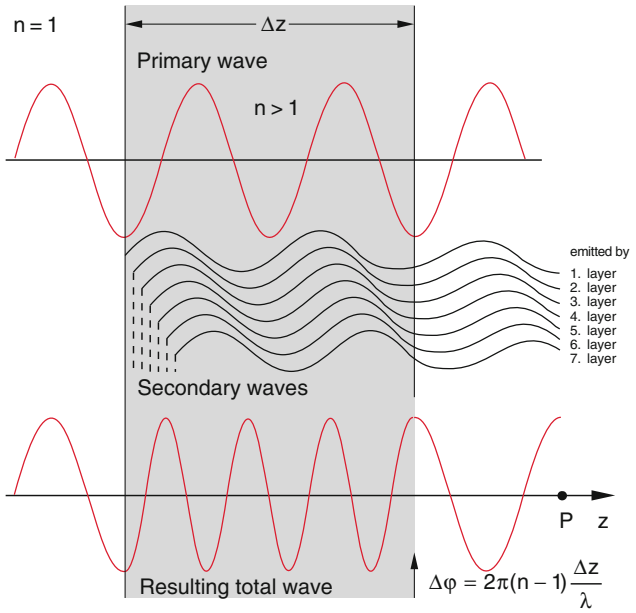


Fig. 8.2 Vivid schematic representation of the delay of a wave during its passage through a dielectric medium. The incident wave is superimposed by secondary waves with a phase delay, emitted by the atoms in the layer of the medium induced to oscillations by the incident wave. The thickness of these layers corresponds to an atomic layer with $\Delta z \approx 0.4$ nm

induced emission is delayed against that of the exciting wave (see Vol. 1, Sect. 11.5).

At the observation point $P(z)$ behind the medium the primary and secondary waves superimpose and form the total field amplitude

$$\boldsymbol{E} = \boldsymbol{E}_\mathrm{e} + \sum_k \boldsymbol{E}_k. \tag{8.2}$$

where the second term represents the sum of the secondary waves emitted by all atoms in the plane z_1 inside the medium.

Because of the phase delay of the secondary waves the total wave arrives at $P(z)$ with a time delay, i.e. it arrives later than without the presence of the medium. Its velocity inside the medium is therefore smaller than in vacuum (Fig. 8.2).

We will at first describe this fact by the broad brush quantity of the refractive index n, before we will derive the relation between n and the atomic characteristics of the medium.

8.1.1 Macroscopic Description

In vacuum the wave would need the time $t = \Delta z/c_0$, to traverse the distance Δz, while inside the medium it takes the time $t_m = \Delta z/c = n\Delta z/c_0$, i.e. it needs the additional time

$$\Delta t = (n-1) \cdot \Delta z/c_0.$$

Behind the medium the wave at the point $P(z)$ is then described by

$$\begin{aligned} \boldsymbol{E}(z) &= \boldsymbol{E}_0 \mathrm{e}^{\mathrm{i}\,\omega[t-(n-1)\Delta z/c - z/c]} \\ &= \boldsymbol{E}_0 \mathrm{e}^{\mathrm{i}\,\omega(t-z/c)} \cdot \mathrm{e}^{-\mathrm{i}\,\omega(n-1)\Delta z/c}. \end{aligned} \tag{8.3}$$

The first factor in (8.3) gives the unperturbed wave which would occur in the absence of the medium.

The influence of the medium is described by the second factor

$$\mathrm{e}^{-\mathrm{i}\Delta\varphi} \quad \text{with} \quad \Delta\varphi = \omega(n-1)\Delta z/c = 2\pi(n-1)\frac{\Delta z}{\lambda}$$

If the phase shift $\Delta\varphi$ is sufficiently small (this is the case for gaseous media with $n - 1 \ll 1$, but is generally not valid for solid media), we can apply the approximation

$$\mathrm{e}^{-\mathrm{i}\varphi} \approx 1 - i\varphi$$

This gives with (8.3) the superposition (8.2) in the simplified form

$$\begin{aligned} \boldsymbol{E}(z) &= \underbrace{\boldsymbol{E}_0 \mathrm{e}^{\mathrm{i}\,\omega\left(t-\frac{z}{c}\right)}}_{\boldsymbol{E}_\mathrm{e}} \underbrace{-\mathrm{i}\,\omega(n-1)\frac{\Delta z}{c}\boldsymbol{E}_0 \mathrm{e}^{\mathrm{i}\,\omega\left(t-\frac{z}{c}\right)}}_{\sum_k \boldsymbol{E}_k} \\ &= \boldsymbol{E}_\mathrm{e} + \boldsymbol{E}_\mathrm{Medium} \end{aligned} \tag{8.4}$$

where the influence of the secondary waves onto the delay of the total wave is described globally be the refractive index n and the thickness Δz of the medium.

8.1.2 Microscopic Model

The second term $\boldsymbol{E}_\mathrm{med}$ in (8.4) can be described by a microscopic but still classical theory. We characterize each atomic electron that is induced by the electromagnetic wave $\boldsymbol{E} = \boldsymbol{E}_0 \cdot \mathrm{e}^{\mathrm{i}(\omega t - kz)}$ to forced oscillations due to the force $\boldsymbol{F} = -e \cdot \boldsymbol{E}$ by the classical model of the damped harmonic oscillator (see Vol. 1, Sect. 11.5).

The oscillation forced by a wave that is linear polarized in x-direction, is described by the equation of motion

$$m\ddot{x} + b\dot{x} + Dx = -eE_0 \mathrm{e}^{\mathrm{i}(\omega t - kz)} \tag{8.5}$$

This gives with the abbreviations $\omega_0^2 = D/m$, $\gamma = b/m$ the oscillation amplitude of the atomic electrons (see 6.43)

$$x_0 = -\frac{eE_0/m}{(\omega_0^2 - \omega^2) + \mathrm{i}\gamma\omega}. \tag{8.6a}$$

Note We have here, contrary to the definition in Vol. 1 defined the damping factor of the amplitude as $\gamma/2$ instead of γ. With this definition γ becomes the damping factor of the power and the following equations then agree with the majority of the literature.

Expanding (8.6a) by $[(\omega_0^2 - \omega^2) - i\gamma\omega]$ yields

$$\begin{aligned} x_0 &= -\frac{(\omega_0^2 - \omega^2 - i\gamma\omega)e/m}{(\omega_0^2 - \omega^2)^2 + (\gamma\omega)^2} E_0 \\ &= -(\alpha + i\beta)E_0 = -\sqrt{\alpha^2 + \beta^2}\, E_0 e^{i\varphi} \qquad (8.6b) \\ \Rightarrow x &= -\frac{e/m}{\sqrt{(\omega_0^2 - \omega^2)^2 + (\gamma\omega)^2}} E_0 e^{i(\omega t + \varphi)}. \end{aligned}$$

This shows that the amplitude of the forced oscillation not only depends on the driving force $-e \cdot E_0$ but also on the frequency difference $\omega_0 - \omega$ and the damping constant γ. The phase shift φ between oscillation amplitude $x(t)$ and the inducing wave $E(t)$ is

$$\tan\varphi = -\frac{\gamma \cdot \omega}{\omega_0^2 - \omega^2} \ . \qquad (8.6c)$$

It also depends on the difference $\omega_0 - \omega$ and the damping factor γ (Fig. 8.3).

The oscillating dipoles with the dipole moment $p(t) = -e \cdot x(t)$ (the valence electrons oscillate against the ion cores which are assumed to be fixed in space) radiate themselves waves (see Sect. 8.6.4). The share of a single dipole measured at time t at the point $P(r)$ with a distance $r \gg x_0$ from the dipole is

$$E_D(r, \vartheta) = -\frac{e\omega^2 x_0 \sin\vartheta}{4\pi\varepsilon_0 c^2 r} e^{i\omega(t - r/c)} \ , \qquad (8.7)$$

where the retardation, i.e. the travelling time $t = r/c$ of the wave from the dipole to the point P has been taken into account.

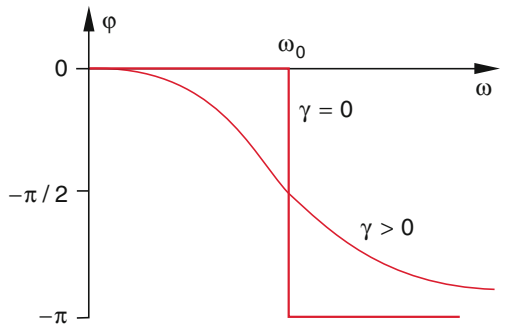


Fig. 8.3 Phase shift φ between oscillation amplitude $x(t)$ of the dipoles and the excitation wave $\boldsymbol{E}(t)$ as a function of the frequency ω of the excitation wave for different values of the damping constant γ

We regard in Fig. 8.4 oscillating dipoles in a thin layer Δz around the plane $z = z_0$. Their number is

$$\Delta N = \Delta z \cdot \int N \cdot dA$$

where N is the density number of dipoles per unit volume and $dA = 2\pi \cdot \rho \cdot d\rho$ the area of the annulus with radius ρ and width $d\rho$ in the x-y-plane.

The total field amplitude generated in the point P by all dipoles in the layer Δz is the superposition of the individual field amplitudes from the single dipoles. Since the distance between the different atoms in the layer is small compared with the wavelength λ of visible light, we can consider the distribution of the dipoles as continuous, We then obtain by integration the total field amplitude of all dipoles in the layer with thickness Δz around $z = 0$

$$E(z) = -\frac{e\omega^2 x_0 e^{i\omega t}}{4\pi\varepsilon_0 c^2} \Delta z \cdot \int_0^\infty N \frac{e^{-i\omega r/c}}{r} \sin\vartheta 2\pi\varrho d\varrho \ . \qquad (8.8a)$$

If the incident light beam has the cross section $\pi \cdot \varrho_{max}{}^2$ only dipoles in the range from $\varrho = 0$ to $\varrho = \varrho_{max}$ are excited. For $z \gg \varrho_{max}$ we can approximate $\alpha \approx 0 \rightarrow \vartheta \approx 90°$ => $\sin\vartheta \approx 1$.

If the density N of oscillating dipoles is constant within the volume $dV = dz \cdot dA$ we can extract N out of the integral and obtain

$$E(z) = -\frac{e\omega^2 x_0 e^{i\omega t}}{2\varepsilon_0 c^2} N \cdot \Delta z \cdot \int \frac{e^{-i\omega r/c}}{r} \varrho d\varrho \ . \qquad (8.8b)$$

With $r^2 = z^2 + \varrho^2 \Rightarrow r dr = \varrho d\varrho$ (in the plane $z = z_0$ is z constant) the integral yields

$$\begin{aligned} \int_0^\infty \frac{e^{-i\omega r/c}}{r} \varrho d\varrho &= \int_{r=z}^\infty e^{-i\omega r/c} dr \\ &= -\frac{c}{i\omega} \left[e^{-i\omega r/c} \right]_z^\infty . \end{aligned} \qquad (8.9)$$

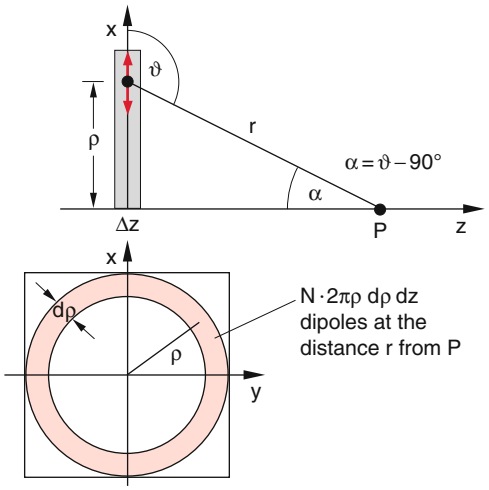


Fig. 8.4 Illustration of the derivation of the electric field E at the point P, caused by dipoles in the layer Δz around $z = 0$

When the diameter of the incident light beam is $d = 2\varrho_{max}$ the range $\varrho > d/2$ does not contribute to the integral because there are no excited dipoles. We therefore can neglect the contribution from dipoles in the range $\varrho_{max} < \varrho < \infty$ corresponding to $\varrho_{max}/\sin\alpha < r < \infty$ to the integral and obtain

$$E(z) = \frac{i\omega e x_0 N}{2\varepsilon_0 c} e^{i\omega(t-z/c)} \cdot \Delta z \,. \tag{8.10}$$

Inserting the expression (8.6a) for the amplitude x_0 the field amplitude generated by $N \cdot \Delta z$ dipoles in a layer Δz

$$E(z) = -i\omega \frac{\Delta z}{c} \cdot \frac{Ne^2}{2\varepsilon_0 m[(\omega_0^2 - \omega^2) + i\omega\gamma]} \cdot E_0 e^{i\omega(t-z/c)} \,. \tag{8.11}$$

Note that the incident light beam diameter $2\varrho_{max}$ does not enter (for $z \gg \Delta z$) into the field amplitude $E(z)$.

This is the additional contribution to the field amplitude, described by the second term in (8.4). The comparison with (8.4) gives for the refractive index the expression

$$n = 1 + \frac{Ne^2}{2\varepsilon_0 m[(\omega_0^2 - \omega^2) + i\omega\gamma]} \,. \tag{8.12a}$$

The refractive index is *a complex number*! We can write this as

$$n = n_r - i \cdot \kappa.$$

It depends on

- the density N of oscillating dipoles, which means the atomic density of the medium
- the frequency difference $\Delta\omega = \omega_0 - \omega$ between the frequency ω of the incident electromagnetic wave and the resonance frequency $\omega_0 = \sqrt{D/m}$ of the oscillating atomic electrons. The latter is determined by the restoring force $(-D \cdot x)$ which is proportional to the displacement from the equilibrium position $x = 0$ and by the electron mass $m = m_e$.

Note The derivation of the refractive index, outlined above, is strictly valid only for optical thin media $(n - 1 \ll 1)$, where the density N of the oscillating dipoles is sufficiently small. This is the case for gases.

Table 8.1 Real part n_r of the complex refractive index for dry air at 20 °C and 1 bar pressure. Here is $n_r \gg \kappa$

λ//nm	$(n-1) \times 10^4$
300	2.915
400	2.825
500	2.790
600	2.770
700	2.758
800	2.750
900	2.745

Example

The refractive index of air at atmospheric pressure is $n = 1.0003$, $(n - 1) \ll 1$ (see Table 8.1).

Remark The approximation in the derivation of (8.12a) has been used twice: Firstly by approximating $e^{-i(n-1)} \approx 1 - i(n - 1)$ when deriving (8.4) from (8.3) and secondly when we assumed that the electromagnetic field emitted by the atomic dipoles should be very small compared to the field of the incident wave. With this approximation we could assume that the amplitude of the exciting wave is equal to that of the incident wave, although we should have used the total amplitude (which depends in the medium on z, because the incident wave is partly absorbed and scattered). For $(n - 1) \ll 1$, which means a small density N of oscillating dipoles, both approximation are well justified. This restriction to media with $n - 1 \ll 1$ is set aside in Sect. 8.3.

8.2 Absorption and Dispersion

We can understand the physical meaning of the complex refractive index, when we write (8.12a) as $n = n_r - i\kappa$. Expanding the fraction in (8.12a) by $[(\omega_0^2 - \omega^2) - i\omega\gamma]$ we get

$$\begin{aligned} n &= 1 + \frac{Ne^2}{2\varepsilon_0 m} \cdot \frac{(\omega_0^2 - \omega^2) - i\omega\gamma}{(\omega_0^2 - \omega^2)^2 + \omega^2\gamma^2} \\ &= n_r - i\kappa. \end{aligned} \tag{8.12b}$$

Inserting this into (8.3), the electric field $\boldsymbol{E}(z)$ of the wave propagating through the medium with thickness Δz becomes with $k_0 = \omega/c$

$$\begin{aligned} \boldsymbol{E}(z) &= \boldsymbol{E}_0 e^{-\omega\kappa\frac{\Delta z}{c}} \cdot e^{-i\omega(n_r-1)\frac{\Delta z}{c}} \cdot e^{i(\omega t - k_0 z)} \\ &= A \cdot B \cdot \boldsymbol{E}_0 \cdot e^{i(\omega t - k_0 z)} \,. \end{aligned} \tag{8.13}$$

The factor $A = e^{-\omega\kappa\Delta z/c}$ describes the attenuation of the amplitude of the wave passing through the medium. After the path length $z = c/(\omega \cdot \kappa)$ the amplitude has decreased to 1/e of its initial value (absorption).

The intensity $I = c \cdot \varepsilon_0 \cdot E^2$ of the wave has then decreased to

$$I(z) = I_0 \cdot e^{-\alpha\Delta z} \tag{8.14}$$

(**Beer's law of absorption**). The quantity

$$\alpha = \frac{4\pi\kappa}{\lambda_0} = 2k_0\kappa \tag{8.15}$$

is the **absorption coefficient**. Its unit is $[\alpha] = 1\,m^{-1}$.

The absorption coefficient α is proportional to the imaginary part of the complex refractive index, where $k_0 = 2\pi/\lambda_0$ is the wavenumber of the wave in vacuum.

The factor $B = e^{-i\omega(n_r-1)\Delta z/c}$ in (8.3) gives the phase shift of the wave while propagating through the medium. This additional phase delay compared with the passage over the distance Δz in vacuum, is

$$\begin{aligned}\Delta\varphi &= \omega(n_r - 1)\Delta z/c \\ &= 2\pi(n_r - 1)\Delta z/\lambda_0\ ,\end{aligned} \tag{8.16}$$

For illustration: the total phase shift of the wave along the distance $\Delta z = \lambda_0$ is $\Delta\varphi_{medium} = n_r \cdot 2\pi$ while in vacuum it is $\Delta\varphi_{vacuum} = 2\pi$. The difference is then $\Delta\varphi = 2\pi \cdot (n_r - 1)$.

Since the wavelength λ is defined as the distance between two phase planes that differ by 2π it follows from the discussion above that the wavelength in the medium is smaller by the factor $1/n_r$ than that in vacuum.

$$\lambda = \frac{\lambda_0}{n_r} \tag{8.17}$$

Because the frequency of the wave does not change in the medium, (see Sect. 8.4) the phase velocity $v_{ph} = \nu \cdot \lambda = (\omega/2\pi) \cdot \lambda$ decreases as

$$v_{Ph} = \frac{c}{n_r}. \tag{8.18}$$

Electro-magnetic waves show in a medium with refractive index $n = n_r - i\kappa$ the wavelength $\lambda = \lambda_0/n_r$ and the phase velocity $v_{ph} = c/n_r$. Their intensity decreases as $I(z) = I_0 \cdot e^{-\alpha z}$ where $\alpha = 2k_0 \cdot \kappa$ is proportional to the imaginary part of the refractive index.

If the medium is characterized by the relative dielectric constant ε (Sect. 1.7.2) and the magnetic induction constant μ (Sect. 3.5.2) the phase velocity is

Table 8.2 Refractive indices $n \approx n_r$ for some optical glasses and transparent materials

λ//nm	480	589	656
FK3	1.470	1.464	1.462
BK7	1.522	1.516	1.514
SF4	1.776	1.755	1.747
SFS1	1.957	1.923	1.910
Quartz glass	1.464	1.458	1.456
Lithiumfluorid LiF	1.395	1.392	1.391
Diamant	2.437	2.417	2.410

$$v_{Ph} = \frac{1}{\sqrt{\varepsilon \cdot \varepsilon_0 \cdot \mu \cdot \mu_0}} = \frac{c}{\sqrt{\varepsilon \cdot \mu}}\ . \tag{8.19}$$

In nonmagnetic media is $\mu \approx 1$. The phase velocity is then

$$v_{Ph} = \frac{c}{\sqrt{\varepsilon}} = \frac{c}{n_r} \Rightarrow n_r = \sqrt{\varepsilon}\ . \tag{8.20}$$

For all transparent media (glass, water, air) the absorption coefficient α for visible light is very small (otherwise the media would not be transparent). The imaginary part κ of the refractive index $n = n_r - i\kappa$ is then small compared to the real part n_r. In this case the refractive index is $n \approx n_r$. Therefore in many equations in optics where mainly transparent media are used (for lenses or prisms) one finds for the refractive index the label n instead of the complex notation $n_r - i\kappa$ (Table 8.2).

From (8.12b) we obtain for the real and the imaginary part of the refractive index n the **dispersion relations**

$$n_r = 1 + \frac{Ne^2}{2\varepsilon_0 m}\frac{(\omega_0^2 - \omega^2)}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}\ , \tag{8.21a}$$

$$\kappa = \frac{Ne^2}{2\varepsilon_0 m}\frac{\gamma\omega}{(\omega_0^2 - \omega^2)^2 + \gamma^2\omega^2}\ , \tag{8.21b}$$

which relate absorption and dispersion of electromagnetic waves in matter with the real and the imaginary part of n (Fig. 8.5).

Note The maximum of the function $\kappa(\omega)$ occurs not exactly at the resonance frequency ω_0, but rather at

$$\omega_{max} = \omega_0 \cdot \left[1 - \frac{\gamma^2}{3\omega_0^2}\right]^{1/2}$$

as can be readily proved by using the condition $d\kappa/d\omega = 0$ in (8.21b) (see Sect. 10.9.2 and Problem 10.14).

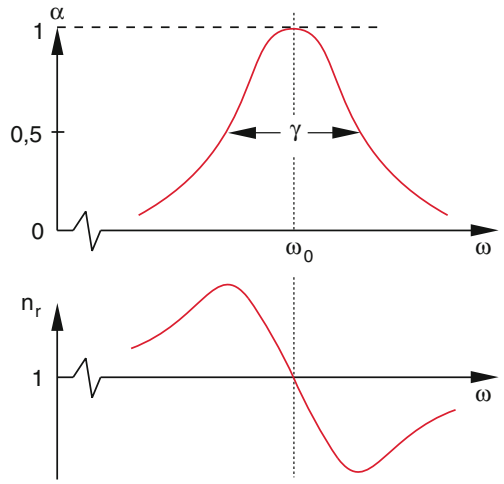


Fig. 8.5 Absorption coefficient $\alpha(\omega) = 2k_0 \cdot \kappa(\omega)$ around an absorption line at $\omega = \omega_0$

Equation (8.12a) for the refractive index anticipated that all damped oscillators had the same resonant frequency ω_0 and the same damping constant γ. In order to apply this classical model to real atoms we have to take into account the following facts:

- The atoms in an absorbing medium have many energy states E_k and can absorb on all frequencies that correspond to transitions between these states, because every transition between two different energy states causes absorption or emission at the frequency ω_k

$$\Delta E = E_k - E_0 = \hbar\omega_k,$$

where $\hbar = h/2\pi$ is the reduced Planck constant (see Vol. 3, Sect. 8.3.1).

- Because an atom with one valence electron can absorb at different frequencies, the probability P_{ik} that it absorbs at a definite frequency ω_{ik} is smaller than the total probability $P = \Sigma P_k$.

For a single absorbing transition the atom has only the fraction ($f_k < 1$) of the absorbency of a classical oscillator. The number $f_k < 1$ is called **oscillator-strength**. It gives the fraction of the absorption probability of a classical oscillator that corresponds to the absorbency of the selected atomic transition. Summing over all possible transitions of the atom the total absorption probability must be that of the classical oscillator. This is equivalent to the condition

$$\sum_k f_k = 1 \qquad (8.22)$$

Sum rule of Thomas, Reiche and Kuhn [1]. The same considerations for the absorption apply for the emission of radiation.

The different excited atoms can absorb energy from the incident wave independently from each other. The total absorption is then the sum of all contributions from the different atoms. The refractive index becomes

$$n = 1 + \frac{e^2}{2\varepsilon_0 m_e} \sum_k \frac{N_k f_k}{(\omega_{0k}^2 - \omega^2) + i\gamma_k \omega} \qquad (8.23)$$

where N_k is the number of atoms per m^3 with the absorption frequency ω_k. Absorption coefficient $\alpha(\omega)$ and refractive index $n_r(\omega)$ have therefore for media with many absorption frequencies a more complex dependence on the frequency ω (Fig. 8.6) as is shown for the single absorption line in Fig. 8.5. In Fig. 8.7 the refractive index $n_r(\omega)$ and the absorption coefficient $\alpha(\omega)$ are illustrated in the vicinity of the two yellow sodium D-lines.

Since in media with dispersion the speed of light $v_{ph}(\omega)$ depends on the frequency ω, phase- and group-velocities differ (see Vol. 1 Sect. 11.9.7). Because $v_{ph} = \omega/k$ we obtain for the group velocity

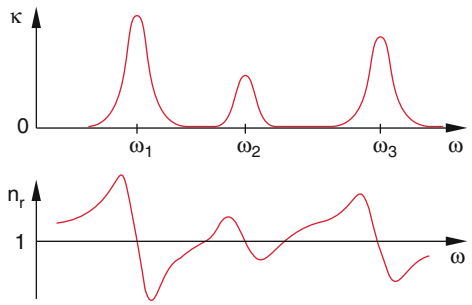


Fig. 8.6 Schematic representation of $\kappa(\omega)$ and $n_r(\omega)$ over a frequency range which includes several absorption frequencies ω_i

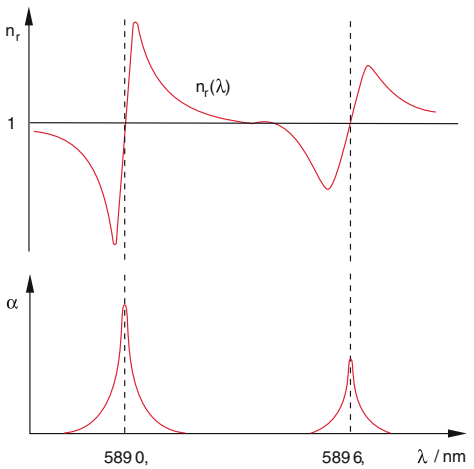


Fig. 8.7 Illustration of dispersion and absorption in the vicinity of the two yellow absorption lines of sodium atoms at $\lambda_1 = 589.0$ nm and $\lambda_2 = 589.6$ nm (without taking account of the hyperfine structure)

$$v_G = \frac{d\omega}{dk} = \frac{d}{dk}(v_{Ph} \cdot k) = v_{Ph} + k \cdot \frac{dv_{Ph}}{dk} .$$

With $k = k_0 \cdot n_r$ and $v_{ph} = c/n_r$ this can be written as

$$v_G = v_{Ph} + k_0 n_r \frac{d}{dk}\left(\frac{c}{n_r}\right)$$
$$= v_{Ph} - k_0 n_r c \frac{1}{n_r^2}\frac{dn_r}{dk}.$$

From the relations

$$k = k_0 \cdot n_r = \frac{\omega}{c} n_r \Rightarrow dk = \frac{n_r}{c} d\omega + \frac{\omega}{c} dn_r$$
$$\Rightarrow \frac{dk}{dn_r} = \frac{n_r}{c}\frac{d\omega}{dn_r} + \frac{\omega}{c}$$
$$\Rightarrow v_G = v_{Ph} - \frac{v_{Ph} k_0}{\frac{1}{v_{Ph}}\frac{d\omega}{dn_r} + k_0} = \frac{v_{Ph}}{1 + \frac{\omega}{n_r}\frac{dn_r}{d\omega}}$$

$$v_G = \frac{c}{n_r + \omega \frac{dn_r}{d\omega}} . \qquad (8.24)$$

This relation gives the following insight:

Figure 8.6 illustrates that there are spectral ranges where the refractive index becomes $n_r < 1$. In these ranges is $v_{ph} = c/n_r > c$ larger than the speed of light in vacuum.

For the determination of the group velocity v_G we calculate $dn_r/d\omega$ from (8.21a).

$$\frac{dn_r}{d\omega} = \frac{Ne^2}{2\varepsilon_0 m} \frac{2\omega\left[(\omega_0^2 - \omega^2)^2 - (\gamma\omega_0)^2\right]}{\left[(\omega_0^2 - \omega^2)^2 + (\gamma\omega)^2\right]^2} . \qquad (8.24a)$$

For $\omega_0^2 - \omega^2 > \gamma\omega_0$ => $dn_r/d\omega > 0$. In this range is $v_G < v_{ph}$ and n_r decreases with increasing wavelength λ. This is called the region of **normal dispersion**. In this range is according to (8.24) always $v_G < v_{ph}$.

The range of **anomalous dispersion** $\Delta\omega_{ad}$, where $dn_r/d\omega < 0$ can be expressed with $\omega_0^2 - \omega^2 = (\omega_0 - \omega)(\omega_0 + \omega) \approx 2\omega_0(\omega_0 - \omega)$ as

$$\omega_0 - \gamma/2 \le \omega \le \omega_0 + \gamma/2 ==> \Delta\omega_{ad} = \gamma.$$

This is the frequency range around an absorption frequency ω_k where, according to (8.21b), the absorption $\kappa(\omega)$ is larger than half of the maximum value $\kappa(\omega_0)$.

In the range of anomalous dispersion the imaginary part $\kappa(\omega)$ of the complex refractive index, and therefore also the absorption coefficient $\alpha(\omega)$ becomes maximum.

The group velocity v_G becomes larger than the speed of light c in vacuum, if the condition

$$n_r + \omega \cdot dn_r/d\omega < 1.$$

is valid. Inserting the expression for n_r from (8.21a) and looking for the derivative $dn_r/d\omega$ one obtains the condition

$$v_G > c \text{ for } |\omega_0 - \omega| < \gamma/2, \qquad (8.24b)$$

This corresponds to the range of anomalous dispersion.

The fact that the group velocity v_G can be larger than the speed of light c is surprising at first sight, because it seems to contradict the special relativity theory, which postulates that the speed of light sets an upper limit for all possible velocities of signal transmission.

The result of (8.24b) is, however, not in contradiction to special relativity theory. This can be seen as follows:

We must distinguish between different velocities

- the phase velocity $v_{ph} = c/n_r$.
- the group velocity $v_G = d\omega/dk$.
- The velocity of energy transport v_E, defined by $v_E = I/w_{em}$ as the ratio of intensity I and energy density w_{em}.
- the velocity of signal transmission.

It turns out that for all media the relation holds

$$\boldsymbol{v_E < c}.$$

In order to transmit a signal the incident light must have a specific intensity profile $I(t)$, as for example a short pulse. The maximum of $I(t)$ can be defined as the signal time. In the spectral range of anomalous dispersion the refractive index $n(\omega)$ changes very fast with ω (see Fig. 8.7). A pulse with width Δt has the frequency width $\Delta\omega \ge 2\pi/\Delta t$. Its frequency width increases with $1/\Delta t$ (Fourier theorem). Because of the large value of $dn/d\omega$ the different frequency components of the pulse have different phase velocities. The superposition of these components after passing through the medium gives a different form of the pulse $I(t)$. Such a deformed pulse $I(t)$ contains the original information only in a modified form. The maximum of the pulse does not travel with the group velocity v_G but with the velocity v_s, which is different from v_G. It turns out that v_s is always smaller than c [2–4].

Remarks

- In recent years specially prepared atoms in a gas which have been optically pumped into so called "dark states" absorb light and reemit it after a particularly long time. The travelling time of light thus becomes extremely long, which corresponds to a velocity of light of a few m/s [5]. However, this apparent low velocity is only due to the long capture time in the atoms and has nothing to do with the speed of light between the capture times, which has still the normal value c. This is analogous to the lower phase velocity $v_{ph} = c/n_r$ in a medium with refractive index

n, where the emission of excited atoms occurs with a time delay (see Sect. 8.1).

- Often the wavenumber k of waves in media is written as a complex number.

$$k = n \cdot \omega/c = n \cdot k_0 = k_0(n_r - i\kappa).$$

The advantage of this convention is that one can use the same expression for waves in vacuum and in matter

$$\begin{aligned} \boldsymbol{E} &= \boldsymbol{E}_0 \cdot e^{i(\omega t - kz)} = \boldsymbol{E}_0 \cdot e^{-\kappa(\omega/c)z} \cdot e^{i(\omega t - n_r(\omega/c)z)} \\ &= \boldsymbol{E}_0 \cdot e^{-(\alpha/2)z} \cdot e^{i(\omega t - n_r k_0 z)}. \end{aligned}$$

8.3 Wave Equation of Electromagnetic Waves in Matter

We start with the Maxwell Eq. (4.26), which have, according to Sects. 1.7.3 and 3.5.2, with the charge density ϱ and the current density $\boldsymbol{j}$ the form

$$\begin{aligned} \nabla \times \boldsymbol{E} &= -\frac{\partial \boldsymbol{B}}{\partial t}, \quad \nabla \cdot \boldsymbol{D} = \varrho\,, \\ \nabla \times \boldsymbol{B} &= \mu\mu_0\left(\boldsymbol{j} + \frac{\partial \boldsymbol{D}}{\partial t}\right), \quad \nabla \cdot \boldsymbol{B} = 0 \end{aligned}$$

where the dielectric displacement density D is defined as

$$\boldsymbol{D} = \varepsilon\varepsilon_0\boldsymbol{E} = \varepsilon_0\boldsymbol{E} + \boldsymbol{P}\,,$$

with the dielectric polarisation $\boldsymbol{P}$.

In the following we will discuss waves in different media based on the wave equation.

8.3.1 Waves in Nonconductive Media

In no conducting media is the current density $\boldsymbol{j} = 0$, because there are no conduction currents. In neutral isolators there are also no free charges, which means $\varrho = 0$.

In a similar way as the derivation of the wave equation in vacuum (Sect. 7.1) we obtain the wave equation for waves in isolators

$$\Delta\boldsymbol{E} = \mu\mu_0\varepsilon\varepsilon_0\frac{\partial^2\boldsymbol{E}}{\partial t^2} = \frac{1}{v_{Ph}^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2} \tag{8.25a}$$

With the phase velocity

$$v_{Ph} = c' = \frac{1}{\sqrt{\mu\mu_0\varepsilon\varepsilon_0}} = \frac{c}{\sqrt{\mu \cdot \varepsilon}}\,. \tag{8.26}$$

An analogous equation

$$\Delta\boldsymbol{B} = \frac{1}{c'^2}\frac{\partial^2\boldsymbol{B}}{\partial t^2} \tag{8.25b}$$

is obtained for the magnetic field,

In non-ferromagnetic media is $\mu \approx 1$ (see Sect. 3.5).

The comparison of (8.26) with (8.1) shows that the refractive index n is related to the dielectric constant ε by

$$n = \sqrt{\varepsilon}\,. \tag{8.26a}$$

Inserting the expression $\boldsymbol{D} = \varepsilon_0\boldsymbol{E} + \boldsymbol{P}$ into the Maxwell equation

$$\mathbf{rot}\boldsymbol{B} = \mu_0\frac{\partial \boldsymbol{D}}{\partial t}$$

We obtain instead of (8.26) the equivalent equation

$$\begin{aligned} \Delta\boldsymbol{E} &= \mu_0\varepsilon_0\frac{\partial^2\boldsymbol{E}}{\partial t^2} + \mu_0\frac{\partial^2\boldsymbol{P}}{\partial t^2} \\ &= \frac{1}{c^2}\frac{\partial^2\boldsymbol{E}}{\partial t^2} + \frac{1}{\varepsilon_0 c^2}\frac{\partial^2\boldsymbol{P}}{\partial t^2}. \end{aligned} \tag{8.25c}$$

This equation contains in concise form the result already discussed in an intuitive way in Sect. 8.1.

The electromagnetic wave in a medium consists of the primary wave, propagating with the vacuum light velocity c (first term in (8.25c)) and the secondary waves generated by the induced atomic dipoles, which superimpose the primary wave with a phase lag (second term in (8.25c)). These secondary waves also propagate through the medium with vacuum light velocity c. The smaller velocity $c' = c/n$ of the total wave is due to the phase shift between secondary and primary waves (Fig. 8.2).

From $\boldsymbol{B} = (\boldsymbol{k} \times \boldsymbol{E})/\omega$ (7.16a) it follows with $\boldsymbol{k} = n \cdot \boldsymbol{k}_0$ and $|\boldsymbol{k}_0| = \omega/c$; $\widehat{k}_0 = \boldsymbol{k}_0/|\boldsymbol{k}_0|$

$$\boldsymbol{B} = \frac{n}{c}(\widehat{\boldsymbol{k}}_0 \times \boldsymbol{E}) = \frac{|n|}{c}(\widehat{\boldsymbol{k}}_0 \times \boldsymbol{E})e^{i\varphi_B}\,, \tag{8.27}$$

where the complex refractive index is written as

$$n = n_r - i\kappa = |n| \cdot e^{i\varphi_B} \text{ with } \tan\varphi = -\kappa/n_r.$$

This shows that in absorbing media ($\kappa \neq 0$) electric field $\boldsymbol{E}$ and magnetic field $\boldsymbol{B}$ are no longer in phase but show a phase shift $\varphi = \arctan(-\kappa/n_r)$.

For the simple case of an isotropic and homogeneous medium and an incident plane wave

$$\mathbf{E} = \{E_x, 0, 0\} = \left\{E_0 \cdot e^{i(\omega t - kz)}, 0, 0\right\}$$

The dielectric polarization P has only one component P_x. For a sufficiently small field E (range of linear optics) it is

$$P_x = N\alpha E_x = N\alpha E_0 e^{i(\omega t - kz)} , \tag{8.28}$$

where N is the number of induced dipoles per m^3 and α is the polarizability (see Sect. 1.7).

Inserting (8.28) into (8.25c) gives

$$\begin{aligned} -k^2 E_x &= -\frac{\omega^2}{c^2} E_x - \frac{\omega^2 N\alpha}{\varepsilon_0 c^2} E_x \\ \Rightarrow\ k^2 &= \frac{\omega^2}{c^2}(1 + N\alpha/\varepsilon_0) . \end{aligned} \tag{8.29}$$

With $v_{ph} = c/nh = \omega/k$ => $n = c\omega/k$ we obtain

$$n^2 = 1 + N\alpha/\varepsilon_0 . \tag{8.30}$$

This equation gives the relation between refractive index n and polarizability α of the atoms in the medium.

The induced dipole moment $p = -e \cdot x$ of each atomic dipole, where the charge $-e$ experiences the displacement x by the electric field $\boldsymbol{E}$ of the electromagnetic wave is then obtained from (8.6a) as

$$p = \frac{e^2 E}{m(\omega_0^2 - \omega^2 + i\gamma\omega)} .$$

With $p = \alpha \cdot E$ the polarizability

$$\alpha = \frac{e^2}{m(\omega_0^2 - \omega^2 + i\gamma\omega)} . \tag{8.31}$$

The comparison with (8.30) finally yields

$$n^2 = 1 + \frac{e^2 N}{\varepsilon_0 m(\omega_0^2 - \omega^2 + i\gamma\omega)} . \tag{8.32}$$

This equation is also valid for large values of n ($n - 1 \gg 1$).

For small values of n ($n - 1 \ll 1$), (8.32) converges with $(n^2 - 1) = (n + 1) \cdot (n - 1) \approx 2(n - 1)$ towards Eq. (8.12a).

Note

(1) The polarizability α and the absorption coefficient α are unfortunately denoted by the same letter, because this is conventional in the physics textbooks.

(2) In magnetic materials the relative magnetic permeability is $\mu \neq 1$. The refractive index then becomes

$$\begin{aligned} n^2 &= \varepsilon \cdot \varepsilon_0 \cdot \mu \cdot \mu_0 = \frac{1}{c^2}\varepsilon\mu \\ \Rightarrow\ n &= \pm\sqrt{\varepsilon\mu} . \end{aligned} \tag{8.33}$$

Recently it has become possible to realize microscopic structures consisting of very small capacitors and inductances where $\varepsilon < 0$ and $\mu < 0$. For such media the minus sign in (8.33) has to be taken. **This means that the refractive index becomes negative!**

This has the consequence that for $\varepsilon < 0$ the pointing vector $\boldsymbol{S} = \varepsilon \cdot \varepsilon_0 c^2 \cdot (\boldsymbol{E} \times \boldsymbol{B})$ points into the opposite direction as the wave vector $\boldsymbol{k}$.

In media with a negative refractive index n the energy flux is opposite to the propagation direction of the wave [6–8] (see also Sect. 8.4.4).

8.3.2 Waves in Conducting Media

When an electromagnetic waves enters a conducting medium with the electric conductivity σ the electric field of the wave induces an electric current with the current density $\boldsymbol{j}$. In the Maxwell Eq. (4.26b) one can no longer set $\boldsymbol{j} = 0$. The derivation of the wave equation proceeds, however, in a similar way as in Sect. (8.3.1). With the relation $\boldsymbol{j} = \sigma \cdot \boldsymbol{E}$ we obtain the wave equation

$$\Delta\boldsymbol{E} = \frac{1}{v_{Ph}^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} + \mu\mu_0\sigma\frac{\partial \boldsymbol{E}}{\partial t} . \tag{8.34}$$

The additional term $\mu\mu_0\sigma \cdot \partial\boldsymbol{E}/\partial t$ corresponds to the damping term $-\gamma \cdot dx/dt$ in the equation of motion of the damped harmonic oscillator (see Vol. 1, Sect. 11.2). The solutions of (8.34) for an incident plane wave propagating in z-direction through the medium is

$$\boldsymbol{E}(z,t) = \boldsymbol{E}_0 \cdot e^{-(\alpha/2)z} \cdot e^{i(\omega t - kz)} \tag{8.35}$$

with the absorption coefficient $\alpha = 2 \cdot k_0 \cdot \kappa$.

We will now discuss the relation between the absorption coefficient α and the electric conductivity σ.

For high frequencies ω of the incident wave the free conduction electrons give the main contribution to the refractive index. Since here the restoring force is zero (contrary to the atomic electrons which are bound to their equilibrium position in atoms by the restoring force with the constant $k = m \cdot \omega^2$) the frequency ω_0 in (8.32) is $\omega_0 = 0$. We then obtain for the refractive index the equation

$$n^2 = 1 - \frac{Ne^2/(\varepsilon_0 m)}{\omega^2 - i\gamma\omega} . \tag{8.36}$$

The damping constant γ is determined by collisions of the free electrons with the lattice atoms. The mean time between two collisions is $\tau = 1/\gamma$.

With (8.36) we define the **plasma frequency**

$$\omega_P = \sqrt{\frac{N \cdot q^2}{\varepsilon_0 \cdot m}} \tag{8.37}$$

The plasma frequency ω_p is the resonance frequency of the free electrons with density N and mass m, which oscillate against the positive charges of the *plasma* (=highly ionized gas) which is altogether electrical neutral (the average positive and negative charges just compensate).

In metals the free electrons with the charge $q = -e$ oscillate under the influence of the electromagnetic wave against the positive charges of the lattice ions. Inserting (8.37) into (8.36) we get

$$n^2 = 1 - \frac{\omega_P^2}{\omega^2 - i\gamma\omega} = 1 - \frac{\omega_P^2}{\omega^2\left(1 - \frac{i}{\omega\tau}\right)} \; . \tag{8.38}$$

With the complex refractive index $n = n_r - i\kappa \Rightarrow n^2 = (n_r - i\kappa)^2 = n_r^2 - \kappa^2 - 2in_r\kappa$ we get after expansion of the fraction in (8.38) with $(1 + i/(\omega\tau))$ and comparing nominator and denominator

$$n_r^2 - \kappa^2 = \frac{1 + \tau^2\left(\omega^2 - \omega_P^2\right)}{1 + \omega^2\tau^2} \tag{8.39a}$$

$$2n_r\kappa = \frac{\omega_P^2\tau}{\omega(1 + \omega^2\tau^2)} \; . \tag{8.39b}$$

In order to determine the electric conductivity $\sigma(\omega)$ as a function of the frequency ω we start with the equation of motion for a damped electron without restoring force under the influence of the electric field $E = E_0 \cdot e^{-i\omega t}$

$$m\left(\frac{dv}{dt} + \gamma v\right) = e \cdot E_0 e^{-i\omega t} \; .$$

With the ansatz $v = v_0 \cdot e^{-i\omega t}$ for the velocity of the electrons we obtain the solution

$$v_0 = \frac{eE_0}{m} \frac{1}{\gamma - i\omega} \; .$$

Since the mean current density j at a charge carrier density N is

$$j = N \cdot e \cdot v_0 = \sigma_{el} \cdot E$$

(see Sect. 2.2) we get

$$\begin{aligned} \sigma_{el} &= \frac{Ne^2}{m} \frac{\tau}{1 - i\omega\tau} = \varepsilon_0\omega_P^2 \frac{\tau}{1 - i\omega\tau} \\ &= \varepsilon_0\omega_P^2 \frac{\tau(1 + i\omega\tau)}{1 + \omega^2\tau^2} \; . \end{aligned} \tag{8.40}$$

The comparison of real- and imaginary parts in (8.40) and (8.39a, 8.39b) gives the relations

$$n_r^2 - \kappa^2 = 1 - \frac{\mathrm{Re}(\sigma_{el})}{\varepsilon_0/\tau} \; ; \quad 2n_r\kappa = \frac{\mathrm{Im}(\sigma_{el})}{\varepsilon_0\omega^2\tau} \tag{8.41}$$

between absorption coefficient $\alpha = 2k_0 \cdot \kappa$ and the electric conductivity σ_{el}.

Note Unfortunately the polarizability α and the absorption coefficient α are denoted in literature by the same letter. Nevertheless, the careful reader will hopefully not be confused.

It is illustrative to consider the two limiting cases of small frequencies ($\omega\tau \ll 1$) and high frequencies ($\omega\tau \gg 1$), where τ is the mean time between collisions of the electrons.

(a) **Low to medium frequencies** ($\omega\tau \ll 1 \ll \omega_p \cdot \tau$)

For this limiting case of low frequencies the electric conductivity σ_{el} can be obtained from (8.40). It is approximately:

$$\sigma_{el} \approx \varepsilon_0 \cdot \tau \cdot \omega_P^2 \; . \tag{8.42a}$$

and is independent of the frequency ω. The complex refractive index at low frequencies is

$$\begin{aligned} n_r - i\kappa &= \sqrt{1 - \frac{\omega_P^2\tau}{\omega(\omega\tau - i)}} \\ &\approx \sqrt{1 - i \cdot \frac{\omega_P^2\tau}{\omega}} \approx \sqrt{-i\frac{\omega_P^2\tau}{\omega}} \; . \end{aligned}$$

With $\sqrt{(-i)} = (1 - i)/\sqrt{2}$ this becomes

$$n_r = \kappa = \sqrt{\omega_P^2\tau/2\omega} \; . \tag{8.42b}$$

Real- and imaginary part of the complex refractive index become equal for low frequencies $\omega\tau \ll 1 \ll \omega_P\tau$!

Example

In a metal is $N \approx 8 \times 10^{28}$ m^{-3} $\Rightarrow \omega_p = 1.6 \times 10^{16}$ s^{-1}. The mean time between collisions of the electrons is $\tau \approx 2 \times 10^{-14}$ s. For the frequency $\omega = 2 \times 10^{13}$ s^{-1} ($\lambda = 94$ µm) we get $\omega\tau = 0.4$, $\omega_p\tau = 320$ and $\omega_p^2\tau = 5 \times 10^{18}$ s^{-1}. This gives

$$n_r = \kappa = 354 \; .$$

The penetration depth of the electromagnetic wave (skin depth) is only

$$\delta = 1/\alpha = c/(2\omega\kappa) = \lambda/4\pi\kappa = 2 \cdot 10^{-8}\,\mathrm{m}.$$

The wave barely penetrates into the metal. The major part is reflected.

(b) **High frequencies**, which are, however, still smaller than the plasma frequency ($\omega_p \cdot \tau > \omega\tau \gg 1$).

In this frequency range the electric conductivity becomes

$$\sigma_{el} \approx i \cdot \varepsilon_0 \cdot \frac{\omega_P^2}{\omega} \qquad (8.42c)$$

and we obtain from (8.38)

$$n^2 \approx 1 - \frac{\omega_P^2}{\omega^2}\,. \qquad (8.42d)$$

For $\omega < \omega_p \Rightarrow n^2 < 0 \Rightarrow n = n_r - i\kappa$ becomes pure imaginary, which implies $n_r = 0$ (Fig. 8.8a). The wave does not propagate in the medium but is totally reflected (Fig. 8.8b). It penetrates, however, over a small distance into the medium (penetration depth).

With the absorption coefficient

$$\alpha = \frac{4\pi\kappa}{\lambda} = \frac{4\pi}{\lambda}\sqrt{\frac{\omega_P^2}{\omega^2} - 1} = \frac{2}{c}\sqrt{\omega_P^2 - \omega^2}\,. \qquad (8.42e)$$

the penetration depth becomes

$$\delta = \frac{1}{\alpha} = \frac{c}{2\sqrt{\omega_P^2 - \omega^2}}\,. \qquad (8.42f)$$

(c) **Very high frequencies** ($\omega > \omega_p$).

For this case the refractive index n becomes real, the imaginary part is zero ($\kappa = 0$).

The medium becomes transparent.

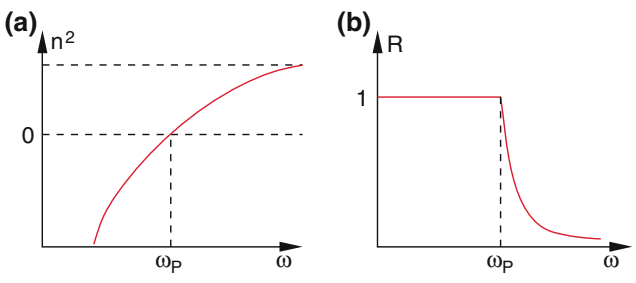


Fig. 8.8 **a**) Square n^2 of the refractive index $n(\omega)$, **b**) Reflectivity R of metals as a function of the frequency ω of the incident wave

Note

(1) In this simple model the plasma frequency ω_p depends solely on the charge carrier density N. The limiting frequency $\omega = \omega_p$, where metals become transparent, is proportional to $\sqrt{N}$.

(2) Te influence of the bound atomic electrons has been neglected, which increases with increasing frequency. Therefore even for $\omega > \omega_p$ a rest absorption remains, which is caused by absorption of the bound atomic electrons [7].

Examples

(a) For copper is $\sigma_{el} \approx 6 \times 10^7$ A/V m, $\tau = 2.7 \times 10^{-14}$ s, $\Rightarrow \omega_p = 1.6 \times 10^{16}\ s^{-1} \Rightarrow \lambda = 120$ nm. This means: For $\lambda > 120$ nm the refractive index of copper is purely imaginary, i.e. copper is highly absorbing. For $\lambda < 120$ nm copper becomes transparent.

(b) For $\omega = 10^{13}\ s^{-1}$ ($\lambda = 180$ µm) is $\omega\tau \ll 1$ and the refractive index becomes $n = 580\,(1 - i) \Rightarrow \alpha = 3.8 \times 10^7\ m^{-1}$. The penetration depth is only $\delta = 1/\alpha \approx 26$ nm.

(c) For $\omega = 3 \times 10^{15}\ s^{-1}$ ($\lambda = 600$ nm) is $\omega\tau \gg 1$ and, according to (8.42d) is $n^2 = -27 \Rightarrow n_r \approx 0$ and $\kappa = 5.2$. The absorption coefficient becomes $\alpha = \approx 10^8\ m^{-1}$. The incident wave is completely reflected and the penetration depth is only $\delta = 10^{-8}$ M = 10 nm.

(d) For $\omega = 3 \times 10^{12}\ s^{-1}$ ($\lambda = 600$ µm) the refractive index becomes $n = 10^7\,(1 - i)$. Real- and imaginary part are equal. The penetration depth becomes $\delta = 15$ nm. However, in this range our simple model fails and the influence of the bound electrons as well as the vibration of the atoms contribute to the absorption.

(e) In the ionized layers of the earth atmosphere (Heaviside layer see Sect. 7.9.4) is $N \approx 10^{11}\ m^{-3} \Rightarrow \omega_p = 2 \times 10^7\ s^{-1} \Rightarrow \upsilon_p = 3$ MHz. Radio waves with $\upsilon < 3$ MHz are totally reflected at the lower side of the Heaviside layer.

8.3.3 The Energy of Electromagnetic Waves in Matter

The wave vector k of a wave in an isotropic medium with refractive index $n = n_r - i\kappa$ is

$$\boldsymbol{k} = n \cdot \boldsymbol{k}_0\,,$$

where $|\boldsymbol{k}_0| = \omega/c$ is the wave vector in vacuum.

The magnetic field of the wave is, according to (8.27),

$$\boldsymbol{B} = \frac{1}{\omega}(\boldsymbol{k} \times \boldsymbol{E}) = \frac{n}{c}(\hat{\boldsymbol{k}}_0 \times \boldsymbol{E})$$
$$= \frac{|n|}{c}(\hat{\boldsymbol{k}}_0 \times \boldsymbol{E})\mathrm{e}^{\mathrm{i}\varphi_B} = \frac{1}{v_{\mathrm{Ph}}}(\hat{\boldsymbol{k}}_0 \times \boldsymbol{E})\mathrm{e}^{\mathrm{i}\varphi_B} .$$

In vacuum $\boldsymbol{B}$ is perpendicular to $\boldsymbol{E}$ and $\boldsymbol{k}$. In matter with a complex refractive index $\boldsymbol{E}$ and $\boldsymbol{B}$ are generally out of phase. If the imaginary part $\mathrm{i}\kappa$ of n is small compared to the real part n_r, the phase shift is negligible.

The Pointing vector of the wave is

$$\boldsymbol{S} = \boldsymbol{E} \times \boldsymbol{H} = \frac{1}{\mu\mu_0}\boldsymbol{E} \times \boldsymbol{B} = \varepsilon\varepsilon_0 v_{\mathrm{Ph}}^2(\boldsymbol{E} \times \boldsymbol{B}) . \qquad (8.43)$$

Inserting the expression (8.27) for $\boldsymbol{B}$ and for $\boldsymbol{E}$ the electric field becomes

$$E = E_0 \cdot \mathrm{e}^{\mathrm{i}\omega(t-nz/c)} = E_0 \cdot \mathrm{e}^{-\frac{\alpha}{2}z} \cdot \mathrm{e}^{+\mathrm{i}\varphi}$$

We obtain for the amount of the Pointing vector

$$|\boldsymbol{S}| = \varepsilon\varepsilon_0 v_{\mathrm{Ph}} E_0^2 \mathrm{e}^{-\alpha z}\cos\varphi\cos(\varphi+\varphi_B) , \qquad (8.44a)$$

where $\alpha = 2k_0 \cdot \kappa$ is the absorption coefficient. The time average of $\langle S\rangle$ can be written as

$$\langle|\boldsymbol{S}|\rangle = \frac{\varepsilon\varepsilon_0 c n_\mathrm{r}}{2|n|^2}\boldsymbol{E}_0^2 \qquad (8.44b)$$

Because

$$\langle\cos\varphi\cdot\cos(\varphi+\varphi_B)\rangle = \langle\cos^2\varphi\cdot\cos\varphi_B - \cos\varphi\cdot\sin\varphi\cdot\sin\varphi_B\rangle$$
$$= \frac{1}{2}\cos\varphi_B$$

and

$$\tan\varphi_B = -\kappa/n_\mathrm{r} \;\Rightarrow\; \cos\varphi_B = \frac{n_\mathrm{r}}{|n|}$$

The time average of the intensity of the wave in a medium with refractive index n is therefore

$$\overline{I} = \frac{1}{2}\varepsilon\varepsilon_0 c n_\mathrm{r}/|n|^2 \cdot E_0^2\mathrm{e}^{-\alpha z}$$
$$= \frac{1}{2}\varepsilon\varepsilon_0 v_{\mathrm{Ph}} E_0^2\mathrm{e}^{-\alpha z}\cos\varphi_B . \qquad (8.44c)$$

8.4 Electromagnetic Waves at the Interface Between Two Media

Assume the incident plane wave

$$\boldsymbol{E}_\mathrm{e} = \boldsymbol{A}_\mathrm{e} \cdot \mathrm{e}^{\mathrm{i}(\omega_\mathrm{e}t - \boldsymbol{k}_\mathrm{e}\cdot\mathbf{r})} \qquad (8.45a)$$

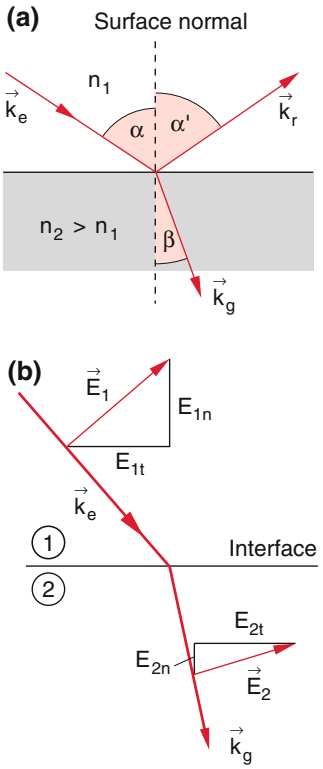


Fig. 8.9 Wave-vectors of incident, reflected and refracted wave at the plane interface between two different media. **b** Dividing the electric field $\boldsymbol{E}$ into the components parallel (tangential) and perpendicular (normal) to the plane of incidence

passes through the interface between two media with different refractive indices n_1 and n_2 (Fig. 8.9). According to the model discussed in Sect. 8.2 the incident wave induces in both media oscillations of the atomic electrons. The question is now how the structure of the total wave can be calculated on both sides of the interface.

The experiment shows that the incident wave splits into a reflected and a refracted part. The refracted wave is

$$\boldsymbol{E}_\mathrm{g} = \boldsymbol{A}_\mathrm{g} \cdot \mathrm{e}^{\mathrm{i}(\omega_\mathrm{g}t - \boldsymbol{k}_\mathrm{g}\cdot\boldsymbol{r})}, \qquad (8.45b)$$

which penetrates into the second medium and has generally another direction as the incident wave.

The reflected wave is

$$\boldsymbol{E}_\mathrm{r} = \boldsymbol{A}_\mathrm{r} \cdot \mathrm{e}^{\mathrm{i}(\omega_\mathrm{r}t - \boldsymbol{k}_\mathrm{r}\cdot\boldsymbol{r})}. \qquad (8.45c)$$

We will now find relations between the amplitudes $\boldsymbol{A}_\mathrm{i}$, the frequencies ω_i and the wave vectors $\boldsymbol{k}_\mathrm{i}$ of the three waves.

8.4.1 Boundary Conditions for Electric and Magnetic Field

We partition the vectors $\boldsymbol{E}$ and $\boldsymbol{B}$ into a part $\boldsymbol{E}_\mathrm{t}$, resp. $\boldsymbol{B}_\mathrm{t}$ parallel to the interface plane (tangential component) and a

part $\boldsymbol{E}_n$ resp. $\boldsymbol{B}_n$ normal (i.e. perpendicular) to the interface (normal component). We then write the field vectors as $\boldsymbol{E} = \boldsymbol{E}_t + \boldsymbol{E}_n$ and $\boldsymbol{B} = \boldsymbol{B}_t + \boldsymbol{B}_n$. This is valid for any arbitrary orientation of $\boldsymbol{E}_i \perp \boldsymbol{k}_i$. For the transition of the wave from medium 1 into medium 2 the tangential component of E and the normal component of B must be continuous, which means $\boldsymbol{E}_t(1) = \boldsymbol{E}_t(2)$ and $\boldsymbol{B}_n(1) = \boldsymbol{B}_n(2)$ (see Sects. 1.7.3 and 3.5.7). We will abbreviate $E_t(1) = E_{1t}$; $E_t(2) = E_{2t}$, etc.

As has been shown in Sect. 1.7 the electric field in a medium with relative permeability ε decreases to $1/\varepsilon$ of its value in vacuum. Since the tangential component does not change at the interface the jump of E must be solely attributed to the normal component. Therefore we get for the *electric field* the relation at the interface between two media with the relative dielectric constants ε_1 and ε_2

$$\frac{E_{1n}}{E_{2n}} = \frac{\varepsilon_2}{\varepsilon_1} = \frac{n_2^2}{n_1^2} \,, \tag{8.46}$$

where we have used the relation $n \approx \sqrt{\varepsilon}$, if absorption and magnetic effects ($\mu = 1$) can be neglected.

For the *magnetic field* the conditions are just opposite: Here is, according to Sect. 3.5.7

$$B_{1n} = B_{2n}; \quad \frac{B_{1t}}{B_{2t}} = \frac{\mu_1}{\mu_2} \,. \tag{8.47}$$

However, since for all non-ferromagnetic materials $\mu \approx 1$ we generally also get $B_{1t} \approx B_{2t}$.

8.4.2 Laws for Reflection and Refraction

We choose our coordinate system such, that the interface plane is the x-z-plane and that the wave vector $\boldsymbol{k}_i$ of the incident wave lies in the x-y-plane (Fig. 8.10). The plane, defined by $\boldsymbol{k}_i$ and the vector $\boldsymbol{N}$ normal to the interface plane is called the **plane of incidence** (in Fig. 8.10 this is the x-y-plane). The continuity of the tangential components then demands

$$E_{et} + E_{rt} = E_{gt} \,. \tag{8.48a}$$

At the origin of the coordinate system ($\boldsymbol{r} = 0$) the insertion of (8.45a, 8.45b) gives

$$A_{et} e^{i(\omega_e t)} + A_{rt} e^{i(\omega_r t)} = A_{gt} e^{i(\omega_g t)} \,. \tag{8.48b}$$

This equation has to be valid for all times. This demands

$$\omega_e \equiv \omega_r \equiv \omega_g, \tag{8.49}$$

All three waves have the same frequency ω.

Since at the transition from medium 1 to medium 2 with different refractive indices n_1 and n_2 the phase velocity

$$v_{ph} = c' = c/n = \nu \cdot \lambda = \omega \cdot \lambda/2\pi$$

changes but the frequency ω does not change, only the wavelength λ must change.

The relation (4.48a), which is valid for all points of the interface, has the consequence that the phases of all three waves must be equal at the interface. We therefore can conclude for all points of the interface:

$$\boldsymbol{k}_e \cdot \boldsymbol{r} = \boldsymbol{k}_r \cdot \boldsymbol{r} = \boldsymbol{k}_g \cdot \boldsymbol{r} \,. \tag{8.50}$$

Since with our choice of the coordinate system the interface is the x-z-plane and the plane of incidence the x-y-plane we get

$$\boldsymbol{k}_r = k_{rx}\hat{\boldsymbol{e}}_x + k_{ry}\hat{\boldsymbol{e}}_y + k_{rz}\hat{\boldsymbol{e}}_z \,,$$
$$\boldsymbol{k}_g = k_{gx}\hat{\boldsymbol{e}}_x + k_{gy}\hat{\boldsymbol{e}}_y + k_{gz}\hat{\boldsymbol{e}}_z.$$

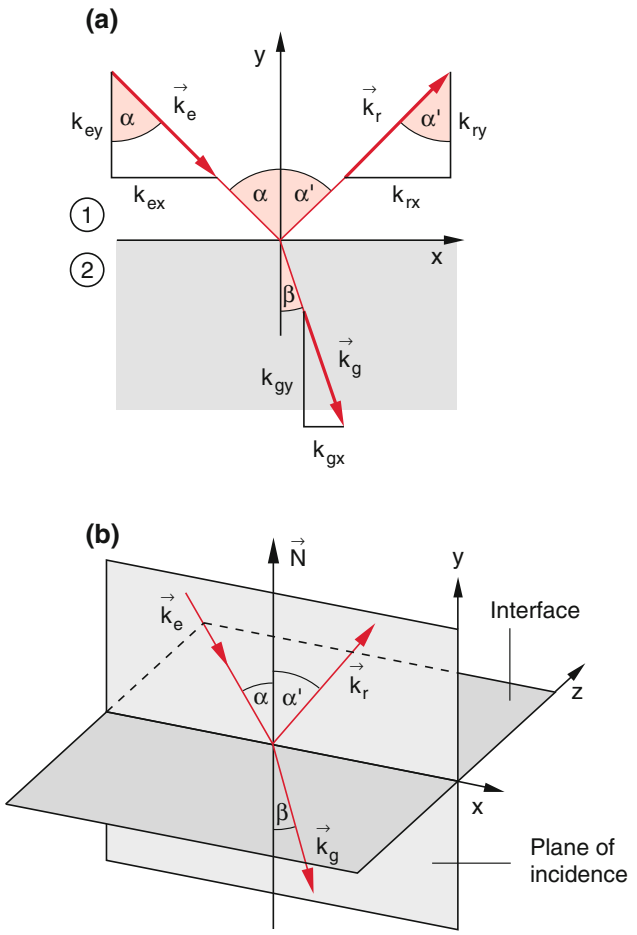


Fig. 8.10 Choice of the coordinate system for the description of reflection and refraction **a**) plane of incidence as drawing plane **b**) perspective representation

Since up to now we do not know the directions of $\boldsymbol{k}_g$ and $\boldsymbol{k}_r$ we make the ansatz

$$\boldsymbol{k}_r = k_{rx}\hat{\boldsymbol{e}}_x + k_{ry}\hat{\boldsymbol{e}}_y + k_{rz}\hat{\boldsymbol{e}}_z ,$$
$$\boldsymbol{k}_g = k_{gx}\hat{\boldsymbol{e}}_x + k_{gy}\hat{\boldsymbol{e}}_y + k_{gz}\hat{\boldsymbol{e}}_z .$$

Inserting this into (8.50) and using (8.51) we obtain

$$k_{ex}x = k_{rx}x + k_{rz}z = k_{gx}x + k_{gz}z . \tag{8.52}$$

This must be valid for all points of the interface, i.e. for arbitrary values of x and z. We therefore get

$$\begin{aligned} k_{ex} &= k_{rx} = k_{gx} , \\ k_{rz} &= k_{gz} = 0 . \end{aligned} \tag{8.53}$$

This means:

> The wave vectors of reflected and refracted wave lie in the same plane as that of the incident wave. All three waves propagate in the plane of incidence.

In Fig. 8.10a the plane of incidence is the drawing plane. From the figure we conclude:

$$\begin{aligned} k_{ex} &= k_e \cdot \sin\alpha, \\ k_{rx} &= k_r \cdot \sin\alpha', \\ k_{gx} &= k_g \cdot \sin\beta. \end{aligned} \tag{8.54}$$

Since the phase velocity of electromagnetic waves is $v_{ph} = c/n$ the amounts of the wave vector is

$$k = \frac{\omega}{c'} = n \cdot \frac{\omega}{c} . \tag{8.55}$$

The frequency ω has the same value in both media. Therefore we get from (8.55) and (8.54)

$$\frac{\sin\alpha}{c_1'} = \frac{\sin\alpha'}{c_1'} = \frac{\sin\beta}{c_2'} . \tag{8.56}$$

This implies:

$$\sin\alpha = \sin\alpha' \;\Rightarrow\; \alpha = \alpha' . \tag{8.57}$$

> The angle of incidence α and the angle of reflection α' are equal. Between the angle of incidence α and the angle of refraction β the relation holds
>
> $$\frac{\sin\alpha}{\sin\beta} = \frac{c_1'}{c_2'} = \frac{n_2}{n_1} \tag{8.58}$$
>
> (**Snell's Law of refraction**).

8.4.3 Amplitude and Polarization of Reflected and Refracted Waves

We will partition the amplitudes of the three waves (8.54) into a part $\boldsymbol{A}_\parallel = \boldsymbol{A}_{ep}$ parallel to the plane of incidence and $\boldsymbol{A}_\perp = \boldsymbol{A}_{es}$ perpendicular to it (Fig. 8.11). This should not be mixed up with the components of the electric field $\boldsymbol{E}_t$ and $\boldsymbol{E}_n$ parallel and perpendicular to the interface plane.

With our choice of the coordinate system the parallel part of the amplitude vector $\boldsymbol{A}_\parallel = \{A_x, A_y, 0\}$ has only an x- and a y-component, while the perpendicular part $\boldsymbol{A}_\perp = \{0, 0, A_z\}$ has only a z-component and is therefore tangential to the interface plane. The continuity of E_n at the interface follows immediately from (8.48b) and (8.49)

$$\boldsymbol{A}_{es} + \boldsymbol{A}_{rs} = \boldsymbol{A}_{gs}. \tag{8.59a}$$

For the tangential component of the magnetic field vector B it follows from (8.47) and (8.27) for non-ferromagnetic materials ($\mu \approx 1$)

$$\boldsymbol{B} = \frac{n}{ck_0}(\boldsymbol{k}_0 \times \boldsymbol{E}) = \frac{n}{\omega}(\boldsymbol{k}_0 \times \boldsymbol{E})$$
$$= 1/\omega(\boldsymbol{k} \times \boldsymbol{E})(\boldsymbol{k}_e \times \boldsymbol{E}_e)_x + (\boldsymbol{k}_r \times \boldsymbol{E}_r)_x = (\boldsymbol{k}_g \times \boldsymbol{E}_g)_x ,$$

This gives the condition for the normal component E_n perpendicular to the plane of incidence

$$k_{ey}A_{es} + k_{ry}A_{rs} = k_{gy}A_{gs} \tag{8.59b}$$

With $k_{ry} = -k_{iy}$ we obtain

$$A_{es} - A_{rs} = \frac{k_{gy}}{k_{ey}}A_{gs} . \tag{8.60}$$

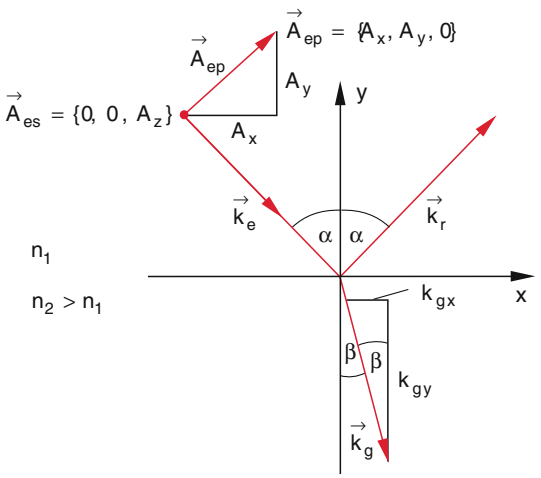


Fig. 8.11 Illustration of the different quantities used in the Fresnel formulas

From (8.59a) and (8.60) it follows

$$A_{gs} = \frac{2}{1+a} A_{es} \quad \text{with} \quad a = k_{gy}/k_{ey} ,$$
$$A_{rs} = \frac{1-a}{1+a} A_{es} .$$

This gives with $k_g = (n_2(n_1)k_{i\vartheta}$

$$\frac{k_{ey}}{k_e} = \cos\alpha; \quad \frac{k_{gy}}{k_g} = \cos\beta .$$

$$a = \frac{n_2 \cos\beta}{n_1 \cos\alpha} .$$

using (8.58) we finally get the amplitude ratios of reflected and refracted waves (**reflection coefficient $\rho_\perp$ and transmission coefficient $\tau_\perp$**)

$$\begin{aligned} \varrho_s &= \frac{A_{rs}}{A_{es}} = \frac{1-a}{1+a} \\ &= \frac{n_1 \cos\alpha - n_2 \cos\beta}{n_1 \cos\alpha + n_2 \cos\beta} \\ &= -\frac{\sin(\alpha-\beta)}{\sin(\alpha+\beta)} , \end{aligned} \tag{8.61a}$$

$$\begin{aligned} \tau_s &= \frac{A_{gs}}{A_{es}} = \frac{2}{1+a} \\ &= \frac{2n_1 \cos\alpha}{n_1 \cos\alpha + n_2 \cos\beta} \\ &= \frac{2\sin\beta\cos\alpha}{\sin(\alpha+\beta)} . \end{aligned} \tag{8.61b}$$

The completely similar derivation for the parallel components yields

$$\begin{aligned} \varrho_s &= \frac{A_{rs}}{A_{es}} = \frac{n_2 \cos\alpha - n_1 \cos\beta}{n_2 \cos\alpha + n_1 \cos\beta} \\ &= -\frac{\tan(\alpha-\beta)}{\tan(\alpha+\beta)} , \end{aligned} \tag{8.62a}$$

$$\begin{aligned} \tau_p &= \frac{A_{gp}}{A_{ep}} = \frac{2n_1 \cos\alpha}{n_2 \cos\alpha + n_1 \cos\beta} \\ &= \frac{2\sin\beta\cos\alpha}{\sin(\alpha+\beta)\cos(\alpha-\beta)} . \end{aligned} \tag{8.62b}$$

Equations (8.61) and (8.62) are the **Fresnel equations**. They form the basis for all calculations of reflection and transmission of electromagnetic waves at the interface between two media with refractive indices n_1 and n_2. Here the incident wave propagates through medium 1 and impinges under the angle α onto the interface. These equations allow one to determine the polarization of reflected and refracted waves for an arbitrary polarization of the incident wave [9].

We will now illustrate the application of the Fresnel formulas by a few examples.

8.4.4 Reflectivity and Transmittance at the Interface

The time average of the intensity $\bar{I}_i$ of a wave incident onto a medium with real refractive index n_1 is according to (8.44c)

$$\bar{I}_e = \varepsilon_0 \varepsilon_1 c_1' \overline{E_e^2} = \frac{1}{2} \varepsilon_0 \varepsilon_1 c_1' A_e^2 \tag{8.63a}$$

With $A_i = (A_\perp + A_\parallel)^{1/2}$ and $c_1' = c_1/n_1$. The reflected mean intensity is then

$$\bar{I}_r = \frac{1}{2} \varepsilon_0 \varepsilon_1 c_1' A_r^2 . \tag{8.63b}$$

The ratio

$$R = \frac{\bar{I}_r}{\bar{I}_e} = \frac{A_r^2}{A_e^2} \tag{8.64a}$$

Is the **reflectivity** of the interface.

Strictly speaking we must consider that a light beam with cross section $\boldsymbol{F}$ incident onto the interface under the angle α covers only the area $\boldsymbol{F}_\alpha = F \cdot \cos\alpha$. The intensity of the incident beam is therefore higher by the factor $1/\cos\alpha$ than on the surface. The correct definition of the reflectivity should be therefore

$$R = \frac{\bar{I}_r \cos\alpha'}{\bar{I}_e \cos\alpha} . \tag{8.64b}$$

However, since the reflection angle $\alpha' = \alpha$ is the same as the incidence angle α, the former definition (8.64a) remains valid.

This is no longer true for the transmission of the refracted beam. Here we have to consider that the refraction angle $\beta \neq \alpha$ differs from the incidence angle α. We therefore define the transmittance T as

$$T = \frac{\bar{I}_t \cos\beta}{\bar{I}_e \cos\alpha} . \tag{8.64c}$$

For I_t we insert the expression

$$\begin{aligned} \bar{I}_t &= \frac{1}{2} \varepsilon_2 \varepsilon_0 c_2' A_g^2 = \frac{1}{2} \cdot \varepsilon_2 \varepsilon_0 \mu_2 \mu_0 c_2'^2 \cdot \frac{1}{\mu_0 c_2'} A_g^2 \\ &= \frac{1}{2} \frac{n_2}{\mu_0 c} A_g^2 , \end{aligned}$$

where we have used the relation $c_2' = 1/(\varepsilon_2 \varepsilon_0 \mu_2 \mu_0)$ and assume that $\mu_2 = 1$. In an analogous way we obtain

$$\bar{I}_e = \frac{1}{2} \frac{n_1}{\mu_0 c} A_e^2 ,$$

This gives for the transmittance

$$T = \frac{n_2 \cos\beta}{n_1 \cos\alpha} \frac{A_g^2}{A_e^2} \,. \tag{8.64d}$$

Since the ratio A_r/A_i differs for the components of A_i parallel or perpendicular to the plane of incidence the reflectivity depends, according to the Fresnel Eqs. (8.61a) and (8.61a) not only on the angle of incidence α and the refractive indices n_1, n_2, but also on the polarization of the incident wave. For the component perpendicular to the plane of incidence we obtain from (8.61a)

$$\begin{aligned} R_s &= \frac{A_{rs}^2}{A_{es}^2} = \left(\frac{n_1 \cos\alpha - n_2 \cos\beta}{n_1 \cos\alpha + n_2 \cos\beta}\right)^2 \\ &= \left(\frac{\sin(\alpha-\beta)}{\sin(\alpha+\beta)}\right)^2 , \end{aligned} \tag{8.65a}$$

while for the parallel component the reflectivity is

$$\begin{aligned} R_p &= \frac{A_{rp}^2}{A_{ep}^2} = \left(\frac{n_2 \cos\alpha - n_1 \cos\beta}{n_2 \cos\alpha + n_1 \cos\beta}\right)^2 \\ &= \left(\frac{\tan(\alpha-\beta)}{\tan(\alpha+\beta)}\right)^2 . \end{aligned} \tag{8.65b}$$

In Fig. 8.12 the reflection coefficient ρ and the reflectivity R are plotted for both components for the case $n_1 < n_2$. For

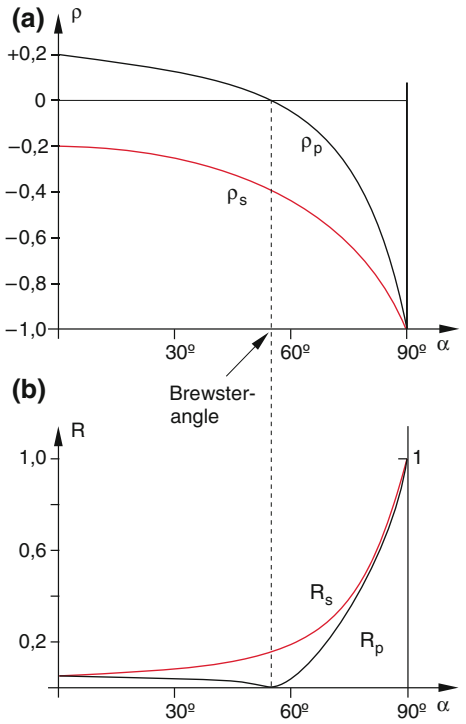


Fig. 8.12 Amplitude reflection coefficient $\rho(\alpha)$ and reflectivity $R(\alpha) = \rho^2(\alpha)$ at an air-glass interface ($n_1 = 1$, $n_2 = 1.5$) for the components with parallel or perpendicular polarization

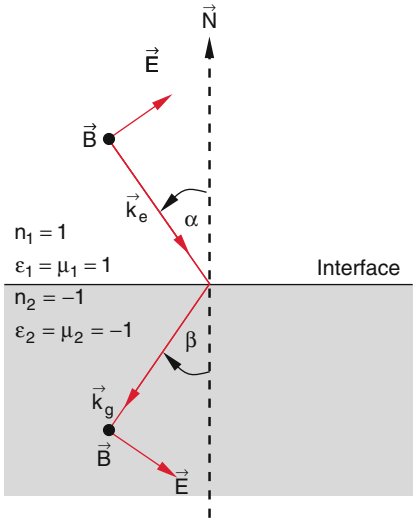


Fig. 8.13 Refraction of an electro-magnetic wave at the interface to a medium with negative refractive index

vertical incidence ($\alpha = 0$) the reflectivity R is equal for both components, as can be deduced already by symmetry arguments. From (8.65a, 8.65b) we get

$$R(\alpha = 0) = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2 . \tag{8.66}$$

Example

The reflectivity at the interface between air and glass ($n_1 = 1$, $n_2 = 1.5$) is for vertical incidence ($\alpha = 0$)

$$R = \left(\frac{0.5}{2.5}\right)^2 = 0.04$$

i.e. 4% of the incident intensity are reflected. The fraction

$$T = \frac{4 n_1 n_2}{(n_1 + n_2)^2} = 0.96$$

is transmitted through the interface into medium 2.

One can readily prove that without absorption for both components the relations hold:

$$T_\parallel + R_\parallel = 1,$$

$$T_\perp + R_\perp = 1,$$

This is valid for any polarization and we can write quite general

$$T + R = 1.$$

Note For materials with negative refractive index (see Sect. 8.4.10) the refracted light beam is on the same side of the interface normal $\boldsymbol{N}$ as the incident beam (Fig. 8.13).

With $n_1 = 1$ and $n_2 < 0$ it follows from (8.58).

$\sin\beta = (n_1/n_2) \cdot \sin\alpha = \sin\alpha/n_2 < 0 \Rightarrow \beta < 0$. This means that the wave vector $\boldsymbol{k}_g$ in Fig. 8.13 points into the direction left of the normal $\boldsymbol{N}$.

8.4.5 Brewster Angle

From (8.62a) it follows the for $\alpha + \beta = 90°$ the amplitude $A_{r\parallel}$ of the reflected beam becomes zero, which means that the reflected wave has no component of the electric field parallel to the plane of incidence (Fig. 8.14), it is completely polarized perpendicular to the plane of incidence.

The angle of incidence $\alpha = \alpha_B$ for which $\alpha + \beta = 90°$ and $R_\parallel = 0$, is called the **Brewster angle**. The wave vectors of reflected and refracted light wave are perpendicular to each other (Fig. 8.14a).

This can be vividly understood as follows:

The incident wave induces the atomic electrons in the interface layer to forced oscillations in the direction of the electric field vector of the transmitted wave (Fig. 8.14b). The amount of the Pointing vector $\boldsymbol{S}$ in the direction ϑ against the dipole axis is proportional to $\sin^2\vartheta$ (see Sect. 8.6.5). The induced dipoles do not radiate into the direction of the dipole axis ($\vartheta = 0$), which is for $\alpha = \alpha_B$ the direction of the reflected beam.

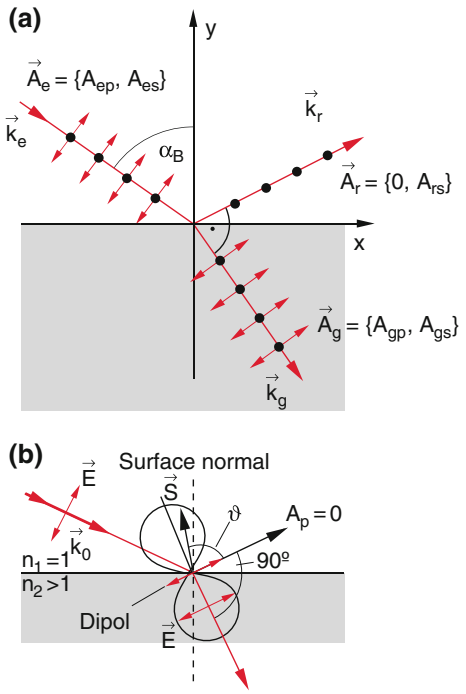


Fig. 8.14 If the incident light forms the Brewster angle α_B with the normal to the interface, the reflected light is linear polarized. **a**) schematic illustration, **b**) physical explanation based on the angular distribution of the intensity radiated by an electric dipole

From $\sin\alpha/\sin\beta = n_2/n_1$ and $\alpha + \beta = 90°$ we obtain the Brewster condition

$$\tan\alpha_B = \frac{n_2}{n_1}. \tag{8.68}$$

Example

For the air-glass interface is $n_1 = 1$ and $n_2 = 1.5$ (for $\lambda = 600$ nm). This gives the Brewster angle $\alpha_B = 56.3°$.

If a linear polarized laser beam with the amplitude vector $A = A_\parallel$ incides onto a glass plate under the angle $\alpha = 56.3°$ the beam passes the plate without any reflection losses, because $A_\perp = 0$. This is used in gas lasers where the discharge tube is sealed on both ends by glass plates under the Brewster angle in order to avoid any reflection losses.

8.4.6 Total Internal Reflection

When a light wave propagates from an optical dense medium 1 into an optically less dense medium 2 ($n_2 < n_1$) one obtains from Snell's law of refraction

$$\sin\alpha = (n_2/n_1)\sin\beta$$

the condition,

$$\sin\alpha_g = n_2/n_1 \tag{8.69}$$

so that the wave can enter medium 2 only for $\sin\alpha < n_2/n_1$ because it is always $\sin\beta \leq 1$ (Fig. 8.15).

For all angles α with $\sin\alpha > n_2/n_1$ the light is reflected at the interface (total reflection). The angle α_c for which $\sin\alpha_c = n_2/n_1$ is the **critical angle of total reflection**.

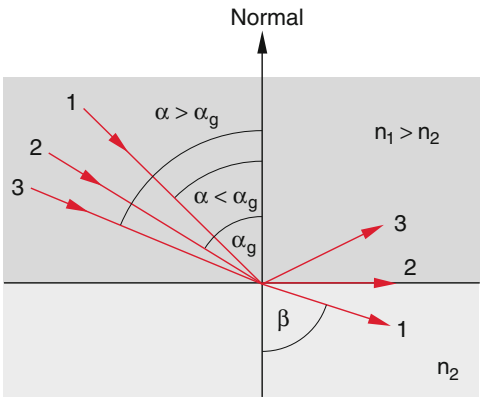


Fig. 8.15 Total reflection of waves that come from the optical dense material and hit the interface under the angle $\alpha > \alpha_g$

Example

For $n_1 = 1.5$ and $n_2 = 1$ the critical angle becomes $\alpha_c = 41.8°$.

The total internal reflection is used in a 90° prism (Fig. 8.16) where the incident light beam hits the interface glass-air under the angle $\alpha = 45°$ and is twice totally reflected. If absorption losses in the prism can be neglected the reflected beam with the opposite direction as the incident beam has the same intensity. The reflectivity of this prism is therefore $R = 1$. Such retroreflectors were placed by the astronauts on the moon. When a laser beam, sent from a telescope on earth hits the retroreflectors, it is reflected back to the earth. Using short laser pulses the measurement of their travel time earth-moon-earth allows the determination of the distance between retroreflector and telescope with an accuracy of a few centimeters.

The total internal reflection is used in optical fibers, which consist of a thin kernel (5–50 µm diameter) with a refractive index n_1, which is sheathed by a cladding with $n_2 < n_1$ (Fig. 8.17). The incident light entering the fiber under an angle $\alpha < 90° - \alpha_c$ against the central line of the fiber is captured within the kernel by total internal reflection and can be thus propagating over long distances.

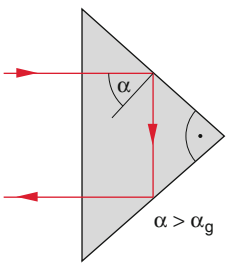


Fig. 8.16 Retro-reflection prism (cats eye)

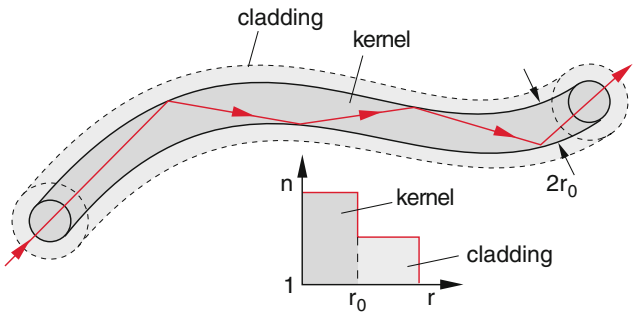


Fig. 8.17 Total reflection in on optical fiber consisting of a quartz core with refractive index n_1 and a cladding with $n_2 < n_1$

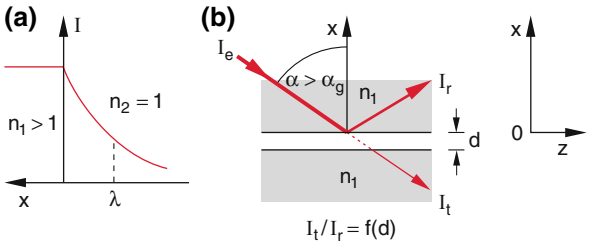


Fig. 8.18 Frustrated total reflection **a**) intensity penetration of fast decreasing light intensity through the interface into the medium with $n_2 < n_1$. **b**) Experimental arrangement for demonstrating frustrated total reflection. With increasing thickness d of the air gap the penetrating intensity strongly decreases

Note

- Total internal reflection only occurs for the transition from the optically dense medium into the optically less dense medium if $\alpha > \alpha_c$ with $\sin \alpha_c = n_2/n_1$.
- Even for total internal reflection the incident wave penetrates within a small layer with thickness $d \approx \lambda$ into the medium with $n_2 < n_1$. This evanescent wave can be detected by the method of **frustrated internal reflection** (Fig. 8.18). If a second glass plate is drawn nearer to the interface glass-air, the light enters the second glass plate if the air gap between the two glass surfaces becomes smaller than the wavelength λ of the light wave.

Such experiments show that the intensity of the wave in the second glass plate decreases exponentially with the thickness Δx of the air gap as $I = I_0 \cdot e^{-\Delta x/\lambda}$. If the medium 2 has no absorption the reflected wave has in spite of its penetration into the second plate the full intensity, i.e. 100% of the incident wave. If, however, an absorbing sample is brought into the air gap, the reflected wave shows the absorption lines of the sample. This technique of absorption spectroscopy of thin absorbing layers based on the evanescent wave, is a very sensitive method.

8.4.7 Change of the Polarization for Inclined Incidence

If linear polarized light falls under the angle α onto an interface the polarization vector of the reflected as well as of the refracted wave is generally turned.

The angle γ_i between the electric field vector $\boldsymbol{E}_i$ of the incident wave and the plane of incidence is defined by

$$\tan \gamma_i = \frac{A_{is}}{A_{ip}}.$$

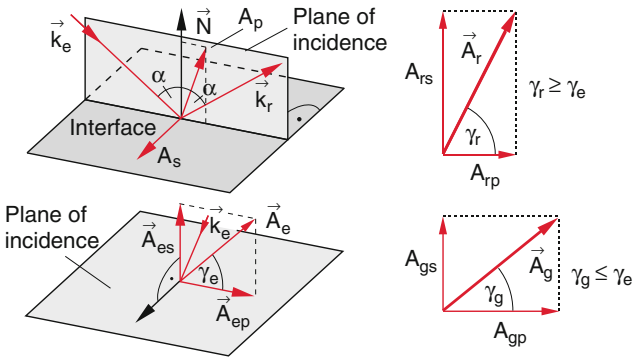


Fig. 8.19 Change of the polarization at reflection

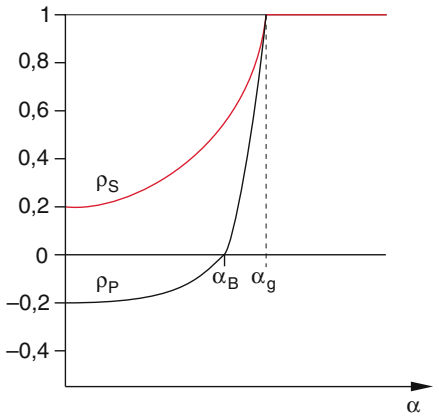


Fig. 8.20 Amplitude reflection coefficient ρ_s and ρ_p at the interface between optical dense and optical thin media

The angle γ_r between the field vector E_r of the reflected wave and the plane of incidence can be obtained from the Fresnel formulas (8.61a and 8.62a) as

$$\tan\gamma_r = \frac{A_{rs}}{A_{rp}} = -\frac{\cos(\alpha-\beta)}{\cos(\alpha+\beta)} \cdot \tan\gamma_e \,. \qquad (8.70)$$

Since $\cos(\alpha-\beta) > \cos(\alpha+\beta)$ it follows (Fig. 8.19):

$$\gamma_r > \gamma_e.$$

The reflection turns the polarization vector $\boldsymbol{E}$ away from the plane of incidence.

Only for vertical incidence ($\alpha = 0°$) and for $\gamma_i = 0°$ or 90° the direction of polarization remains unchanged.

For the refracted wave the angle γ_g is obtained from

$$\tan\gamma_g = \frac{A_{gs}}{A_{gp}} = \cos(\alpha-\beta) \cdot \tan\gamma_e \,. \qquad (8.71)$$

Since $\cos(\alpha-\beta) \leq 1$ it follows $\gamma_g \leq \gamma_e$.

The refraction turns the polarization vector $\boldsymbol{E}$ towards the plane of incidence.

8.4.8 Phase Shift at the Reflection

In the following we consider only media with no absorption ($\kappa = 0$). If the wave is reflected at the interface to an optically dense medium 2 ($n_2 > n_1$) we conclude from (8.61a) with $\cos\beta > \cos\alpha$ that for the reflected wave the amplitude $\boldsymbol{A}_{r\perp}$ of the component perpendicular to the plane of incidence changes sign against the component $\boldsymbol{A}_{i\perp}$ of the incident wave. This means:

Under reflection at the optically dense medium the component perpendicular to the plane of incidence suffers a phase jump of π.

For the component $A_p = A_\parallel$ in Fig. 8.11 a phase jump of π has occurred, if the y-component changes sign.

From (8.62a) we can conclude that the reflection coefficient ϱ_p becomes negative for $(\alpha+\beta) > \pi/2$.

Since $\alpha+\beta = \pi/2$ is the condition for the Brewster angle α_B, where the amplitude of the reflected wave is zero, the parallel component suffers a phase jump only for $\alpha > \alpha_B$ for the reflection at the optically dense medium. The transition from $\alpha < \alpha_B$ to $\alpha > \alpha_B$ is not discontinuous because $A_\parallel(\alpha_B) = 0$ (Fig. 8.20).

In Fig. 8.21 the phase jump $\Delta\varphi$ under reflection at the interface between air and glass is shown for both components as the function of the angle α of incidence.

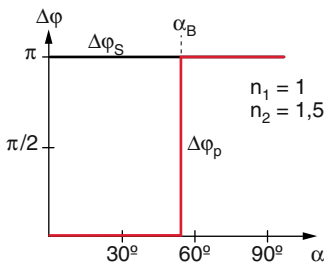


Fig. 8.21 Phase jump at the reflection at the optical dense medium, shown for the example of the interface air-glass

Note For vertical incidence ($\alpha = 0°$) the distinction between $A_\perp$ and $A_\parallel$ becomes meaningless, since all planes through the direction of the incident light beam are planes of incidence. If for $\alpha > 0$ the plane of incidence is defined it follows for $\alpha \to 0$ from (8.61a and 8.61b) for the two components with $n_2 > n_1$

$$A_{r\perp}/A_{i\perp} = A_{r\parallel}/A_{i\parallel} = (n_1 - n_2)/(n_1 + n_2) < 0$$

This means that both components suffer a phase jump. One can therefore make the statement: For $n_2 > n_1$ (interface air-glass) the wave suffers at the reflection a phase jump of π.

Under reflection at the optically thin medium ($n_2 < n_1$) => $\alpha < \beta$) it can be concluded from (8.61a) that the component $A_\perp$ does not suffer a phase jump. For the parallel component $A_\parallel$ (8.62a) shows that for $(\alpha + ß) < \pi/2$ i.e. for $\alpha < \alpha_B$ a phase jump of π occurs. For $a_B \leq \alpha \leq \alpha_c$ (critical angle of total reflection) the phase jump is zero, for $\alpha > \alpha_c$ it increases from $\Delta\varphi = 0$ up to $\Delta\varphi = 90°$ (Fig. 8.22).

For the *refracted* wave the phase jump is zero for all cases discussed above.

For total reflection the phase jumps differ for the two components (Fig. 8.22). This can be seen when (8.61a, 8.61b) and (8.62a, 8.62b) are rewritten using (8.69). For example, reducing the fraction (8.61a, 8.61b) by n_1 yields

$$\varrho_s = \frac{\cos\alpha - \sqrt{\sin^2\alpha_g - \sin^2\alpha}}{\cos\alpha - \sqrt{\sin^2\alpha_g - \sin^2\alpha}}. \qquad (8.72)$$

For $\alpha > \alpha_g$ the radicand becomes negative and numerator and denominator both become complex. However, recalculating shows that $\varrho \cdot \varrho^* = 1 \to$ the reflectivity is $R = 1$.

Under total internal reflection ($n_2 < n_1$) one obtains for the for the component $A_{r\perp}$ the phase jump $\Delta\varphi_\perp$ defined by

$$\tan\left(\frac{\Delta\varphi_s}{2}\right) = \frac{1}{\cos\alpha}\sqrt{\sin^2\alpha - \left(\frac{n_2}{n_1}\right)^2} \qquad (8.73a)$$

and for the component $A_{r\parallel}$

$$\tan\left(\frac{\Delta\varphi_p}{2}\right) = \frac{n_1^2}{n_2^2\cos\alpha}\sqrt{\sin^2\alpha - \left(\frac{n_2}{n_1}\right)^2}. \qquad (8.73b)$$

More detailed information can be found in [7, 8].

8.4.9 Reflection at Metal Surfaces

Metals absorb electromagnetic waves within a wide frequency range.

The imaginary part κ of the refractive index $n = n_r - i\kappa$ is in the visible range generally larger than the real part n_r (see Sect. 8.3.2)

For the determination of the reflectivity of an interface air-metal using the Fresnel formulas (8.61a, 8.61b) we have to insert $n_1 = 1$ and $n_2 = n_r - i\kappa$. This gives for a real amplitude of the incident wave complex expressions for the amplitudes $A_{r\parallel}$ and $A_{r\perp}$ of the reflected wave. This implies that the amplitude as well as the phase suffer changes under reflection at a metal surface.

The phase jumps $\Delta\varphi$ between reflected and incident wave are given by

$$\tan(\Delta\varphi) = -\frac{b}{a} = \frac{\mathrm{Im}(z)}{\mathrm{Re}(z)} = \frac{2\kappa}{1 - n^2 - \kappa^2}.$$

They can take values between 0 and π and are generally different for $A_{r\parallel}$ and $A_{r\perp}$ (see Problem 8.5). Therefore the polarization state of the reflected wave differs from that of

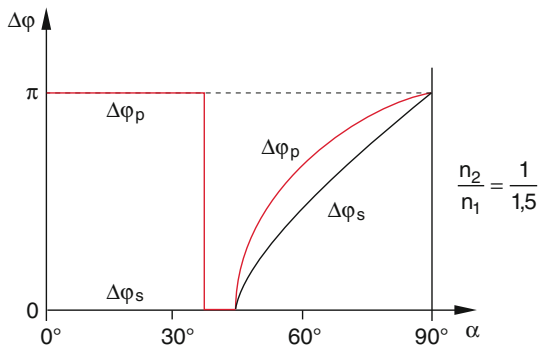


Fig. 8.22 Phase jumps for A_p and A_s for $n_1 > n_2$ and at total reflection for $n_1 = 1.5\ n_2$ and for different incidence angles α

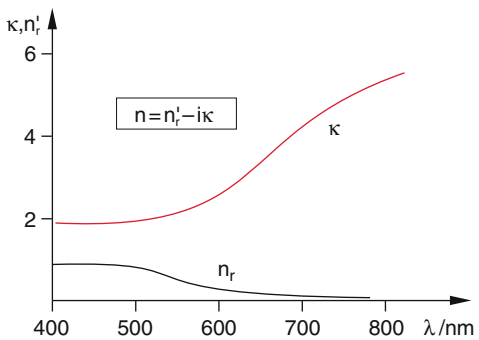


Fig. 8.23 Wavelength-dependent real- and imaginary part of the refractive index of gold

the incident wave. An exception is linear polarized light with the electric vector $\boldsymbol{E}$ parallel or perpendicular to the plane of incidence. For all other directions of $\boldsymbol{E}$ the reflection at metal surfaces generates elliptical polarized light.

For vertical incidence ($\alpha = 0°$) we obtain from (8.66) with $n_1 = 1$ and $n_2 = n_r - i\kappa$ the reflectivity

$$R = |\varrho_s|^2 = \left|\frac{n_r - i\kappa - 1}{n_r - i\kappa + 1}\right|^2 = \frac{(n_r - 1)^2 + \kappa^2}{(n_r + 1)^2 + \kappa^2} \qquad (8.74)$$

It depends on the real part as well as the imaginary part of the complex refractive index. Since both parts depend, according to (8.39a), on the frequency ω and therefore on the wavelength λ (Fig. 8.23) the reflectivity R becomes wavelength-dependent.

Equation (8.74) shows that for $\kappa \gg n_r$ the reflectivity approaches $R \approx 1$.

Example

The refractive index of aluminum at $\lambda = 600$ nm is: $n_r = 0.95$ and $\kappa = 6.4$. The reflectivity at vertical incidence is then $R = 0.91$.

This demonstrates: The surface of strongly absorbing materials has a high reflectivity (see Table 8.3).

The transmission of a thin absorbing layer with the thickness Δz is

$$T = \frac{I_t}{I_e} = e^{-\alpha \Delta z} = e^{-4\pi\kappa\Delta z/\lambda_0}$$

The absorption coefficient $\alpha(\lambda)$ and therefore also $\kappa(\lambda) = \alpha(\lambda)/2k_0$ depend on the wavelength λ (see Fig. 8.6). Those wavelengths λ_m where $\kappa(\lambda)$ becomes maximum are preferentially reflected (see 8.74).

The surfaces of strongly absorbing materials where the refractive index shows a jump, have a reflection coefficient that is proportional to the absorption coefficient.

Table 8.3 Real part n_r and imaginary part κ of the complex refractive index $n = n_r - i\kappa$ and reflectivity R of some metals between $\lambda = 500$ and 1000 nm

Wavelength in nm	Metal	n_r	κ	R
500	Copper	1.031	2.78	0.65
500	Silver	0.17	2.94	0.93
500	Gold	0.84	1.84	0.50
1000	Copper	0.147	6.93	0.99
1000	Silver	0.13	6.83	0.99
1000	Gold	0.18	6.04	0.98

Note If the absorption does not change suddenly but smoothly over a small distance of a wavelengths the reflectivity converges to zero and the incident radiation is completely absorbed. Examples are surfaces covered with soot, black velvet or the surface of the sun.

Experiment:

When writing with a red transparency marker onto a transparency the writing appears under illumination by white light red in transmission but green in reflection, because the red writing absorbs green (Fig. 8.24). The reflected light can be best seen, if the transparency is placed on a dark background and is illuminated from above.

Remark In Fig. 8.24 the incident white light falls onto the surface under an inclined angle $\alpha \neq 0°$ different from the assumption in (8.74), because in the experiment the reflected light should be separated from the incident light. This does not change, however, the conclusions drawn from (8.74).

8.4.10 Media with Negative Refractive Index

In the foregoing sections of this chapter we have discussed that the phase velocity of electromagnetic waves in matter with the constants μ and ε is given by the relation

$$v_{ph} = (\varepsilon \cdot \varepsilon_0 \cdot \mu \cdot \mu_0)^{-1/2} = c/n_r$$

where $n_r = \sqrt{\varepsilon}\mu$ is the real part of the refractive index. For several years researchers succeeded in producing special materials with periodic structures for which both μ and ε are negative, as long as the wavelength of the radiation is larger than the lattice constant a of these periodic structures (a is

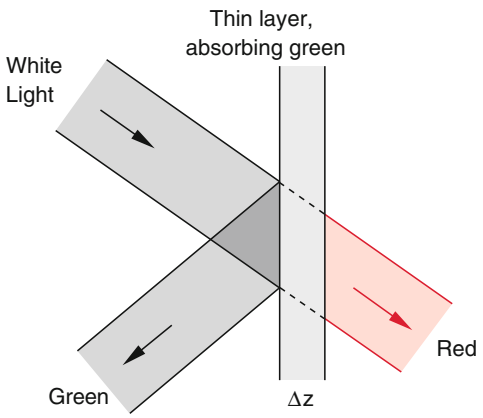


Fig. 8.24 Experimental demonstration that reflectivity and absorbance of strongly absorbing materials are proportional

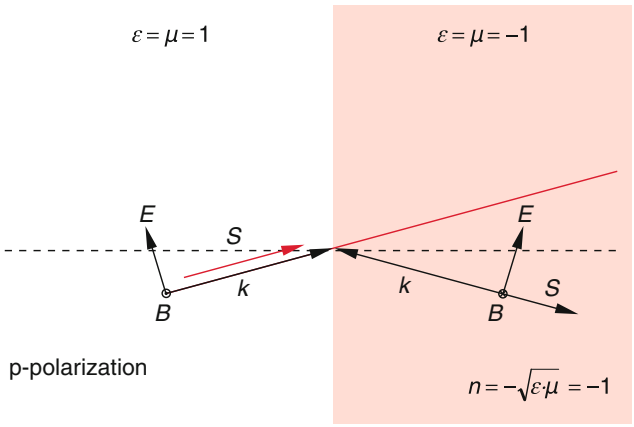


Fig. 8.25 Propagation of an electro-magnetic wave in a "normal" medium with $\varepsilon = \mu = 1$ (left part) and in a "meta-material" with $\varepsilon = \mu = -1$ (right part). Note that in the metamaterial the wave-vector $\boldsymbol{k}$ is antiparallel to the pointing vector $\boldsymbol{S}$

the length of the unit cell of the periodic structure). It turns out that the definition of the refractive index has to be broadened to (Fig. 8.25)

$$n_r = \pm\sqrt{\varepsilon \cdot \mu}$$

where both signs are possible, while before we have only used the + sign. In such with special techniques produced "*meta-materials*" with periodic structures $a < \lambda$ the minus sign has to be used and the refractive index becomes negative. The propagation of light in these meta-materials differs from that in ordinary media with $n > 0$ [9].

The periodic structures are produced by vapor deposition of many micro-resonators, which are arranged in a periodic way. They consist of tiny L-C-circuits which are realized by small squares with silver walls (dimensions w < λ, thickness c) and a hole with width g (Fig. 8.26a). The relative

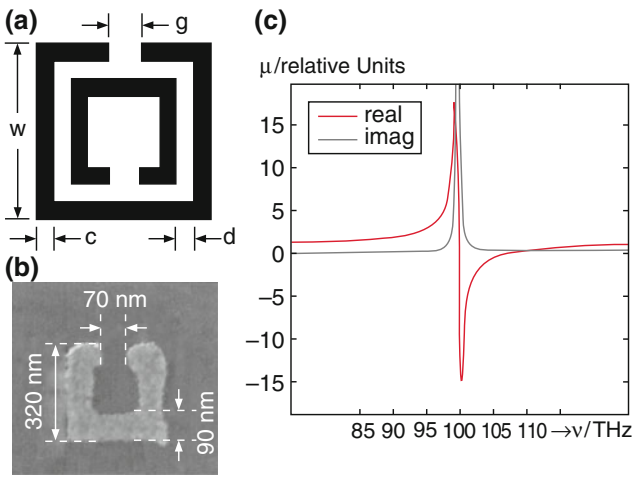


Fig. 8.26 **a**) Schematic representation of a miniaturized LC circuit. **b**) Experimental realization **c**) resonance curve for real and imaginary part of the relative magnetic permeability μ [9, 10, 19]

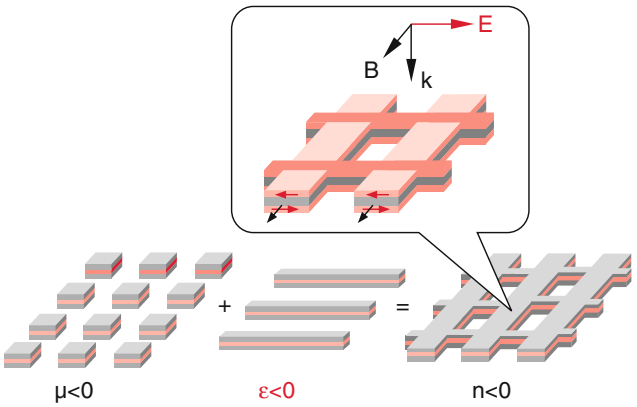


Fig. 8.27 Realization of meta-materials with $n_r < 0$ as superposition of micro-resonators ($\mu < 0$) as layered structures and periodical arrangement of conductive thin rods [9, 10, 19] and Zhang et al. Opt. Express 13, 4922 (2005)

permeability μ around the resonance frequency, shown in Fig. 8.26c, illustrates that just above the resonance frequency μ becomes negative.

Meanwhile it is possible to further minimize the periodic structures and to achieve even in the visible range at λ = 760 nm a negative refractive index $n_r = -0.6$. These devices, shown in Fig. 8.27, are fabricated by a photo-lithographic technique. They consist of many micro-resonators in the form of layer structures causing $\mu < 0$ and a sequence of electrical conductive micro-rods which cause a negative electric permeability $\varepsilon < 0$.

Pendry [10] has proved that with lenses of such meta-materials light can be focused much better than with conventional lenses and that the limitation by diffraction (see Sect. 11.3) can be essentially reduced [8].

8.4.11 Photonic Crystals

Photonic crystals are periodic configurations of dielectric media with alternative different dielectric constants ε and therefore different refractive indices (Fig. 8.28). They can be

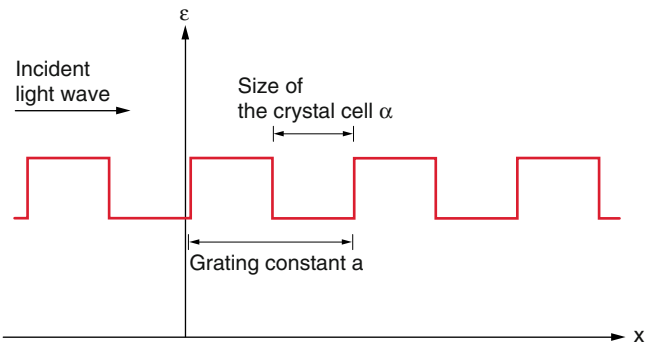


Fig. 8.28 Model of a one-dimensional optical crystal with periodic spatial variation of the dielectric constant ε

realized by transparent materials (glass, semiconductors etc.) into which periodic structures (e.g. periodic micro-cylinders) are implanted with a period a which is about equal to the wavelength of the incident light wave. These structures have a lasting effect on the propagation of light through the photonic crystal [10, 11].

If a light wave falls onto the photonic crystal part of its amplitude is reflected at the interfaces of the periodic spatial arrangement of the microstructure. For $a = m\lambda$ the parts reflected at successive interfaces superimposes constructively because their phase difference is 2 m · π. The total reflectivity of the crystal is increased.

For a = (2m − 1) · λ/4 their superposition is destructively (phase difference (2m − 1) π) and the reflected light is diminished or even completely suppressed.

In Fig. 8.28 is the spatial distribution of ε shown for the model of a one-dimensional photonic crystal.

8.5 Light Propagation in Anisotropic Media; Birefringence

In anisotropic media the restoring force $\boldsymbol{F_r} = -k_r \cdot \boldsymbol{r}$ (in the model of the oscillating dipole), which bonds the oscillating atomic electron to its equilibrium position depends on the direction of the oscillation in the crystal. This implies that the resonance frequencies $\omega_i = (k_{ri}/m)^{1/2}$ of the absorption lines differ for the different polarization directions of the incident electromagnetic wave. According to (8.32) this has the consequence, that the refractive index n depends not only on the frequency ω but also on the direction of the $\boldsymbol{E}$-vector and the $\boldsymbol{k}$-vector of the wave, i.e. on the direction of polarization and propagation of the wave (Fig. 8.29).

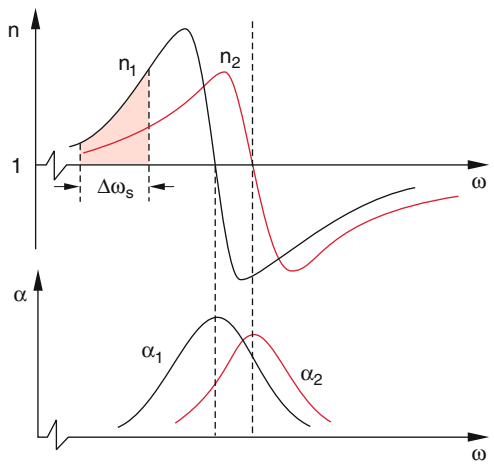


Fig. 8.29 Refractive indices $n_1(\omega)$ and $n_2(\omega)$ and absorption coefficient α in the vicinity of n absorption line n for two mutually perpendicular polarizations of a wave propagating in an anisotropic crystal. The visible range is marked by the red shaded area with spectral width $\Delta\omega_s$

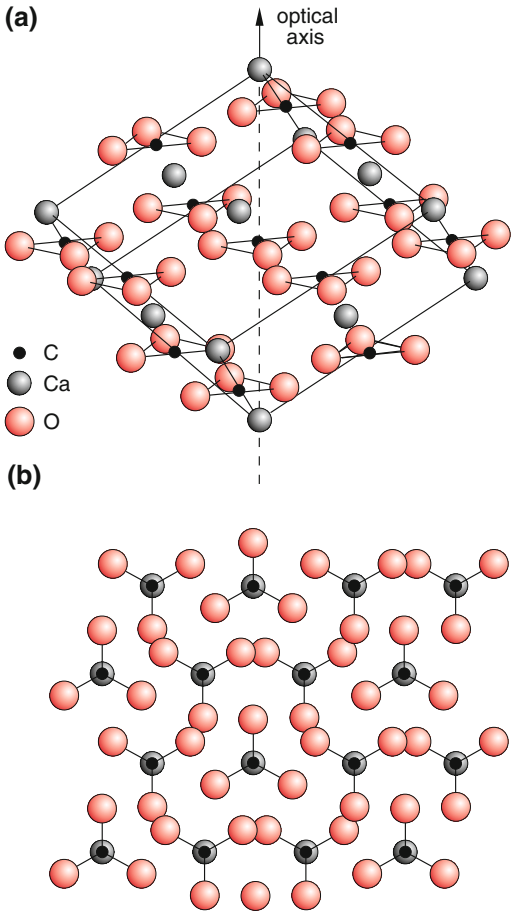


Fig. 8.30 Crystal structure of $CaCO_3$. **a**) Spatial arrangement of the atoms **b**) cut through a $CaCO_3$-crystal perpendicular to the optical axis

The optical anisotropy depends on the crystal structure. In Fig. 8.30 the spatial configuration of the atoms in a calcite crystal $CaCO_3$ is illustrated. One can see, that there exists a preferential direction (perpendicular to the drawing plane), called the **optical axis**. The atomic configuration is, however, not rotationally symmetric around the optical axis. This illustrates that the restoring forces onto the atomic electrons depend on the direction in the drawing plane in Fig. 8.30b, due to the anisotropic force fields caused by the positively charged ions.

Note The optical axis is no geometrical line but indicates that propagation direction in the crystal where all polarization directions have the same refractive index.

8.5.1 Propagation of Light Waves in Anisotropic Media

A simple mechanical experiment can illustrate the conditions for the propagation of light waves in anisotropic media: Two

spiral springs with different force constants k_r are acting on a mass m in the x-resp. the y-direction (Fig. 8.31). In the equilibrium position m rests in the point P(0, 0). With a thread connected to the mass m a force

$$\mathbf{F} = \{F_x,\ F_y\}$$

is exerted on m which points into the direction of the taut thread. The mass, however, does not follow the direction of the threat but moves into the direction $\mathbf{\Delta s} = \{k_x \cdot x, k_y \cdot y\}$. For each point of the trajectory Δs the total force acting on m (traction force F_t plus restoring force $\mathbf{F}_r = -\{k_x \cdot x, k_y \cdot y\}$ is zero.

For our oscillator model of the propagation of electromagnetic waves through anisotropic media this means: The direction of oscillation of the induced dipoles is not necessarily parallel to the inducing electric field vector **E** of the incident wave. The mathematical formulation of this situation is the description of the relative dielectric constant ε by a tensor instead by a scalar as in isotropic media. This tensor is written in form of a matrix

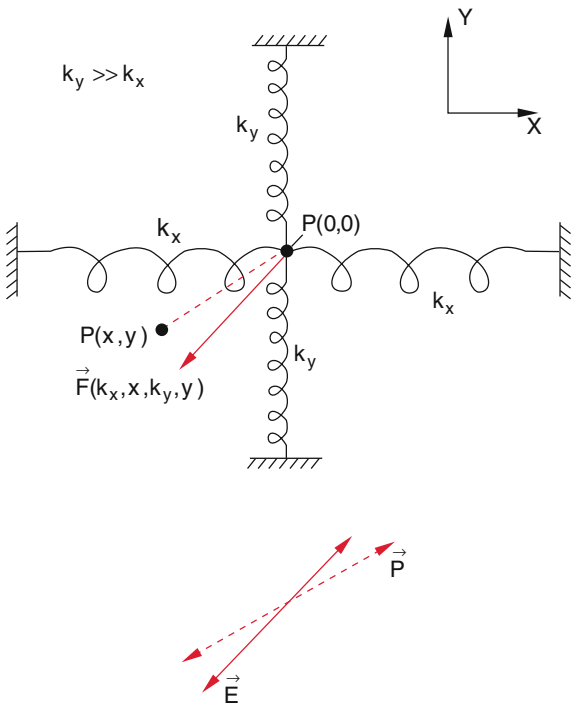


Fig. 8.31 Mechanical model for illustration of optical birefringence. The directions of the acting force **a**) and the elongation are not parallel for unequal restoring forces **b**) for the optical case this means: Inducing field and polarization do not point into the same direction

$$\tilde{\varepsilon} = \begin{pmatrix} \varepsilon_{xx} & \varepsilon_{xy} & \varepsilon_{xz} \\ \varepsilon_{yx} & \varepsilon_{yy} & \varepsilon_{yz} \\ \varepsilon_{zx} & \varepsilon_{zy} & \varepsilon_{zz} \end{pmatrix}. \tag{8.75}$$

The relation between electric field amplitude $\boldsymbol{E}$ and dielectric displacement density $\boldsymbol{D}$ is then, instead of (1.64) for isotropic media given by

$$\boldsymbol{D} = \tilde{\varepsilon} \cdot \varepsilon_0 \boldsymbol{E} \tag{8.76a}$$

where $\tilde{\varepsilon}$ is a tensor with the nine components $\varepsilon_{i,k}$ which can be written by the three equations for the components as

$$\begin{aligned} \frac{1}{\varepsilon_0} D_x &= \varepsilon_{xx} E_x + \varepsilon_{xy} E_y + \varepsilon_{xz} E_z, \\ \frac{1}{\varepsilon_0} D_x &= \varepsilon_{yx} E_x + \varepsilon_{yy} E_y + \varepsilon_{yz} E_z, \\ \frac{1}{\varepsilon_0} D_x &= \varepsilon_{zx} E_x + \varepsilon_{zy} E_y + \varepsilon_{zz} E_z. \end{aligned} \tag{8.76b}$$

Note, that $\boldsymbol{E}$ and $\boldsymbol{D}$ are generally no longer parallel as in isotropic media. When the dielectric displacement density $\boldsymbol{D}$ is expressed by

$$\boldsymbol{D} = \varepsilon_0 \boldsymbol{E} + \boldsymbol{P}, \tag{8.76c}$$

one can write the dielectric polarization $\boldsymbol{P}$ as

$$\boldsymbol{P} = \boldsymbol{D} - \varepsilon_0 \boldsymbol{E} = \varepsilon_0 \left(\tilde{\varepsilon} - \tilde{1}\right) \cdot \boldsymbol{E} = \varepsilon_0 \cdot \tilde{\chi} \cdot \boldsymbol{E}. \tag{8.76d}$$

In anisotropic media the polarization $\boldsymbol{P}$ is generally no longer parallel to the electric field $\boldsymbol{E}$ and the direction of the oscillation of the induced dipoles is then also not parallel to the acting force $\mathbf{F} = q \cdot \mathbf{E}$, (Fig. 8.31b) analog to our mechanical model in Fig. 8.31a. The susceptibility $\tilde{\chi} = (\tilde{\varepsilon} - \tilde{1})$ is a two-stage tensor and (8.76d) can be written as an equation for the components

$$\boldsymbol{P}_i = \varepsilon_0 \cdot \sum_{j=1}^{3} \chi_{ij} \boldsymbol{E}_j \quad (i = x, y, z), \tag{8.76e}$$

This shows that each component P_i of the dielectric polarization can depend on all three components E_j of the incident wave.

In order to investigate the propagation of an electromagnetic wave in insulating charge-free ($\rho = 0$) anisotropic media, we use the two Maxwell equations

$$\operatorname{div} \boldsymbol{D} = 0 \text{ and } \operatorname{div} \boldsymbol{B} = 0.$$

From these equation follow the relations

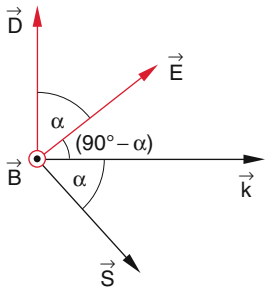


Fig. 8.32 When a light wave propagates in an anisotropic crystal the vectors $\boldsymbol{k}$; $\boldsymbol{E}$; $\boldsymbol{D}$ and $\boldsymbol{S}$ all lie in the same plane perpendicular to $\boldsymbol{B}$, but $\boldsymbol{E}$ is no longer perpendicular to $\boldsymbol{k}$

$$\boldsymbol{D} \cdot \boldsymbol{k} = 0, \; \boldsymbol{B} \cdot \boldsymbol{k} = 0. \tag{8.77}$$

Both vectors $\boldsymbol{D}$ and $\boldsymbol{B}$ are perpendicular to the wave vector $\boldsymbol{k}$. With the relation (8.27)

$$\boldsymbol{B} = (n/c) \cdot (\boldsymbol{k}_0 \times \boldsymbol{E})$$

we conclude that $\boldsymbol{B} \perp \boldsymbol{E}$.

From the definition (7.21) of the Pointing vector

$$\boldsymbol{S} = \varepsilon_0 c^2 (\boldsymbol{E} \times \boldsymbol{B}) \quad (\text{for } \mu = 1)$$

it follows that $\boldsymbol{B} \perp \boldsymbol{S}$. Since in dielectric media no electric current flows, is **rot** $\boldsymbol{B} = \mu\mu_0 \cdot \partial \boldsymbol{D}/\partial t$ (see Sect. 8.3). This implies $\boldsymbol{B} \perp \boldsymbol{D}$.

The consequence of the relations above is the following:

Since $\boldsymbol{B}$ is perpendicular to $\boldsymbol{k}$, $\boldsymbol{E}$, $\boldsymbol{D}$ and $\boldsymbol{S}$, the latter four vectors all have to lie in a plane $\perp \boldsymbol{B}$ (Fig. 8.32). $\boldsymbol{E}$ and $\boldsymbol{D}$ include the angle α, which is determined by the tensor (8.75). An important consequence is that the direction of the wave vector $\boldsymbol{k}$ is no longer identical with the direction of the energy flow $\boldsymbol{S}$. The two vectors $\boldsymbol{k}$ and $\boldsymbol{S}$ include the same angel α as $\boldsymbol{E}$ and $\boldsymbol{D}$ because $\boldsymbol{E} \perp \boldsymbol{S}$ and $\boldsymbol{D} \perp \boldsymbol{k}$. While the phase planes are perpendicular to $\boldsymbol{k}$ the energy flows in the direction of $\boldsymbol{S}$.

In anisotropic crystals the direction of light propagation and energy flow are generally different.

The electric field vector $\boldsymbol{E}$ is perpendicular to $\boldsymbol{S}$ but not to $\boldsymbol{k}$. The wave is no longer strictly transversal. The electric field $\boldsymbol{E}$ has a component in the direction of $\boldsymbol{k}$.

8.5.2 Refractive Index Ellipsoid

In non-absorbing media the tensor elements ε_{ik} in (8.75) are real numbers and for media without optical activity the tensor becomes symmetric, i.e. $\varepsilon_{ik} = \varepsilon_{ki}$. In this case the number of tensor components reduces to six. One can always choose a coordinate system (x, y, z) where the coordinate axes are oriented in such a way that all non-diagonal elements of (8.75) are zero and the tensor becomes diagonal (*principal axis transformation*)

$$\tilde{\varepsilon}_{\text{pa}} = \begin{pmatrix} \varepsilon_1 & 0 & 0 \\ 0 & \varepsilon_2 & 0 \\ 0 & 0 & \varepsilon_3 \end{pmatrix} \tag{8.78}$$

The principal values ε_1, ε_2, ε_3 are obtained by the diagonalization of the corresponding matrix (8.75). These principal values of ε correspond to three values of the refractive index n

$$n_1 = \sqrt{\varepsilon_1}, \quad n_2 = \sqrt{\varepsilon_2}, \quad n_3 = \sqrt{\varepsilon_3}.$$

If the vector

$$n = \{n_1, n_2, n_3\}$$

is plotted in a coordinate system with axes n_1, n_2, n_3 (principal axis system), its endpoint describes the ellipsoid

$$\frac{n_x^2}{n_1^2} + \frac{n_y^2}{n_2^2} + \frac{n_z^2}{n_3^2} = 1 \tag{8.79}$$

which is called the **index ellipsoid** (Fig. 8.33). The lengths of the principal axes of this ellipsoid give the principal values n_i of the refractive index.

Crystals for which $n_1 = n_2 \neq n_3$ are called *optical uniaxial crystals.*

Their index ellipsoid shows rotational symmetry about the z-axis; which corresponds to the crystallographic c-axis of the uniaxial crystal. For $n_3 > n_1 = n_2$ the crystal is called optical positive, for $n_3 < n_1 = n_2$ it is optical negative.

If a plane wave falls into the direction of the wave vector $\boldsymbol{k}$ onto an uniaxial crystal the plane perpendicular to $\boldsymbol{k}$ cuts

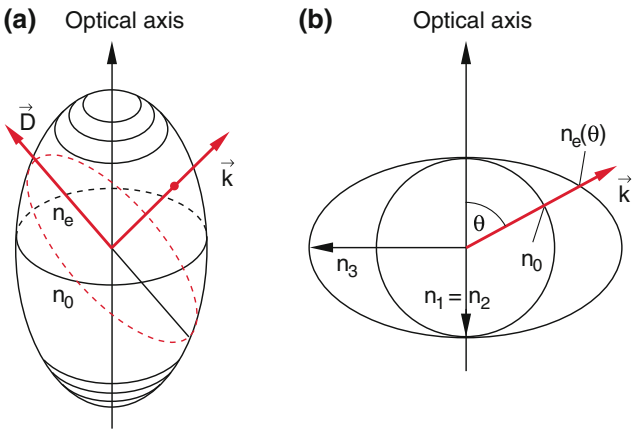


Fig. 8.33 **a**) Rotationally symmetric index ellipsoid with the symmetry axis in the direction of the optical axis. **b**) two-dimensional representation of the extraordinary refractive index $n_e(\theta)$ and the ordinary index n_0 which is independent of θ for a positive uniaxial crystal

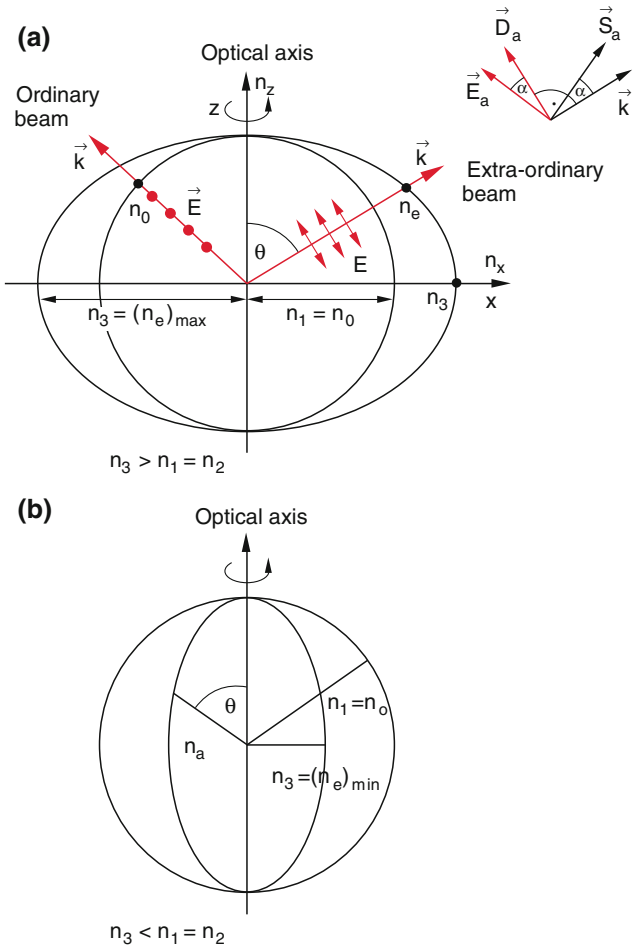


Fig. 8.34 Cut through the index ellipsoid **a**) for a positive and **b**) for a negative uniaxial optical crystal. The distance between the center and the intersection of the propagation direction and the circle resp, ellipse give the refractive indices n_0 resp. n_e for this propagation direction. The red arrows represent two waves in arbitrary directions where only the ordinary beam or the extraordinary beam is shown. The ellipsoids arte rotationally symmetric about the optical axis

the ellipsoid in an ellipse (Figs. 8.33 and 8.34). The vector $\boldsymbol{D}$ lies in this plane. The length of the line segment in the direction of $\boldsymbol{D}$ from the origin to the ellipse gives the refractive index n for this wave and therefore also its phase velocity $v_{ph} = c/n$. It exists a special direction of $\boldsymbol{k}$ for which the slice plane is a circle. This direction is the **optical axis** of the crystal. For this direction of $\boldsymbol{k}$ the refractive index does not depend on the orientation of $\boldsymbol{D}$.

For the general case $n_1 \neq n_2 \neq n_3 \neq n_1$ there are two directions of $\boldsymbol{k}$ for which the slice plane is a circle. In such biaxial crystal there are two optical axes. For all waves propagating into the direction of one of these optical axis the refractive index and the phase velocity of the waves are independent of the direction of $\boldsymbol{E}$. In this case $\boldsymbol{E}$ and $\boldsymbol{D}$ point into the same direction.

Table 8.4 Ordinary refractive indices $n_o = n_1$ and extraordinary index $n_a(90°) = n_3$ for some birefringent uniaxial optical crystals at $\lambda = 589.3$nm

Crystal	n_o	n_a
Crystal quartz	1.5443	1.5534
Calcite	1.6584	1.4864
Tourmaline	1.669	1.638
ADP = Ammonium-Dihydrogen-Phosphate	1.5244	1.4791
KDP = Potassium Diphosphate	1.5095	1.4683
Cadmium sulfid	2.508	2.526

For an uniaxial crystal the z-axis is chosen as the optical axis. Then the x-z-plane through the origin cuts the index ellipsoid, which is rotationally symmetric about the z-axis, in an ellipse for one polarization component ($\boldsymbol{E}$ in the x-z-plane), while the cut for the other component in the x-y plane gives a circle (Fig. 8.34). The refractive index for the component in the x-y-plane does not depend on the angle θ between optical axis and wave vector $\boldsymbol{k}$. It behaves as in an isotropic medium and is called the ordinary refractive index, while the refractive index for the other component in the x-z-plane, which does depend on the angle θ is the extraordinary refractive index. Its maximum value in positive uniaxial crystals is obtained for $\theta = 90°$, where $n_e = n_3$ and $\boldsymbol{k} \parallel \boldsymbol{x}$ (Fig. 8.34a) its minimum value $n_e = n_0$ for $\theta = 0°$ ($\boldsymbol{k} \parallel \boldsymbol{z}$). For our choice of the coordinate axes the light waves with $\boldsymbol{E} = \{0, E_y, 0\}$ are ordinary waves while those with $\boldsymbol{E} = \{E_x, 0, E_z\}$ are extraordinary waves. Table 8.4 compiles some values of n_0 and n_e for some uniaxial crystals.

For crystals with lower symmetry there is no longer a preferential axis and the propagation of light in such crystals is much more complex. There exists no ordinary wave with a refractive index that is independent of the direction of k but there are two extraordinary waves with refractive indices that depend on the direction of k. For optical biaxial crystals with two optical axes there are three refractive indices $n_1 \neq n_2 \neq n_3 \neq n$. The index ellipsoid has no longer rotational symmetry. With the vector $\boldsymbol{r} = \{x, y, z\} = 1/\sqrt{(\rho_{em} \cdot \varepsilon_0)} \cdot \{D_x, D_y, D_z\}$, where $D = \{D_x, D_y, D_z\}$ is the displacement density vector, and $n_i^2 = \varepsilon_i$ ($i = x, y, z$) one obtains the index ellipsoid

$$1 = \frac{x^2}{n_x^2} + \frac{y^2}{n_y^2} + \frac{z^2}{n_z^2}.$$

8.5.3 Birefringence

When a parallel unpolarized light beam enters a calcite crystal ($CaCO_3$) it splits into two beams with different polarization (Fig. 8.35). One beam follows Snell's law of refraction (8.58). For $\alpha = 0$ is also $\beta = 0$. It is therefore

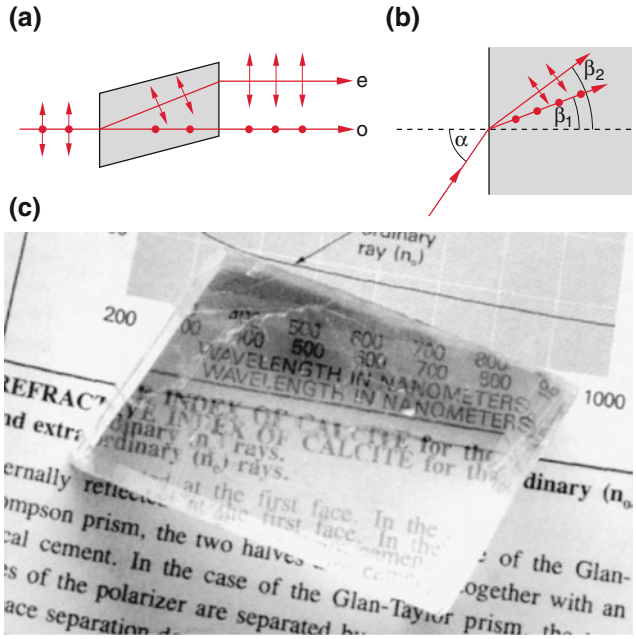


Fig. 8.35 Optical birefringence. **a**) vertical incidence **b**) inclined incidence **c**) Illustration of birefringence in a calcite crystal. The incident unpolarized light is split into ordinary and extraordinary beam which are linearly polarized orthogonal to each other

called the **ordinary beam** (see last section). The second beam has even for A = 0 a refraction angel$\beta \neq 0$ (**extra-ordinary beam**).

The two beams are polarized orthogonal to each other. The ordinary beam is polarized perpendicular to the optical axis of the crystal, while the **E**-vector of the extra-ordinary beam is parallel to the optical axis. Crystals that split the incident light beam into two components, are called *birefringent crystals*.

As has been discussed in Vol. 1, Sect. 11.11 the refraction can be understood with Huygens principle. The propagation direction is the normal to the envelope of the wave fronts of the elementary waves, which are emitted from each point hit by the primary wave (Fig. 8.36).

If the incident light impinges perpendicular to the optical axis onto the crystal (Fig. 8.36a) the phase velocity of the wave does not depend on the propagation direction in the crystal. This is true for both polarization directions. The phase surfaces for each elementary wave (in Fig. 8.36a is only one indicated) originating from the point A are spheres and their cuts with the *x*-*y*-plane are circles. However, the phase velocities differ for the ordinary and the extraordinary wave, because n_o is different from n_e. the tangent from the point B to the phase velocity circles gives the phase front of the total wave for ordinary and extraordinary wave. The wave vectors k_o and k_e are perpendicular to these tangents. They point into different directions.

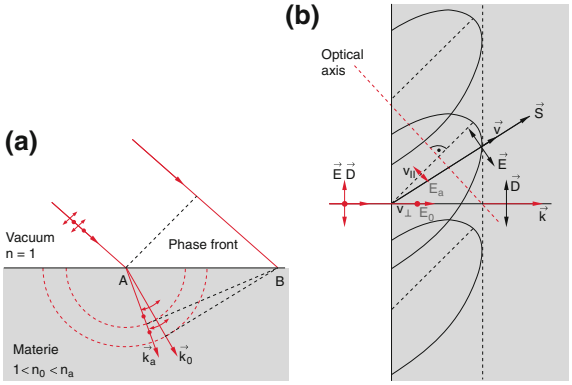


Fig. 8.36 **a**) Birefringence of the incident light, if the optical axis is perpendicular to the drawing plane. **b**) Elliptical wave front for the extraordinary wave with Pointing vector ***S*** and wave vector ***k***

If the direction of the incident light is inclined against the optical axis (in Fig. 8.36b lies the optical axis in the drawing plane) the phase velocity of the extraordinary wave (polarization vector is parallel to the optical axis) depends on the inclination angle against the optical axis (Fig. 8.34). Therefore the cuts of the phase surfaces with the *x*-*y*-plane are ellipses, for the ordinary wave (polarization direction perpendicular to the optical axis) they are circles.

If the wave impinges vertically onto the surface of the crystal (Fig. 8.35a) the ordinary wave passes through the crystal into the same direction as the incident wave (there is no refraction, i.e. the refractive angle $\beta = \alpha$ is equal to the angle of incidence). The extraordinary wave is refracted ($\beta \neq \alpha$),

In Fig. 8.37 the generation of the elliptical phase surface is again illustrated for the general case. The ***E***-vector of the

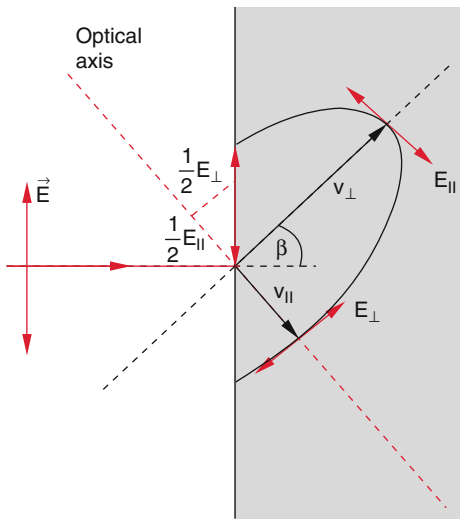


Fig. 8.37 Origin of the elliptical wave fronts in a birefringent crystal when the polarization plane of the incident wave forms an arbitrary angle against the optical axis

incident wave is divided into the component $E_{\parallel}$ parallel and a component $E\perp$ perpendicular to the optical axis. The phase velocities for the two polarization directions are $v_{\parallel}$ and $v_{\perp}$. The refraction angle β for the extraordinary component is then $\sin\beta = v_{\parallel}/c$.

The magnitude of the splitting depends on the angle α of the incident beams against the optical axis and on the difference between the refractive indices n_o and n_e.

Because of the different phase velocities and the different wave vectors $\boldsymbol{k}$ for the two polarization directions wave vector $\boldsymbol{k}$ and Pointing vector $\boldsymbol{S}$ have in birefringent crystals generally different directions. In Fig. 8.38 the directions of field vector $\boldsymbol{E}$, displacement density vector $\boldsymbol{D}$, wave vector $\boldsymbol{k}$ and Pointing vector $\boldsymbol{S}$ are illustrated for the general case in a birefringent crystal.

The tangent to the phase surfaces of the different elementary waves gives the phase surface of the total wave. The propagation vector (wave vector) $\boldsymbol{k}$ is perpendicular to this phase plane. The direction of the Pointing vector $\boldsymbol{S}$ gives the direction of the energy flux. It is

$$\boldsymbol{S} \sim \boldsymbol{E}_{\parallel}^2 \boldsymbol{v}_{\parallel} + \boldsymbol{E}_{\perp}^2 \boldsymbol{v}_{\perp}$$

Since $\boldsymbol{S} = \boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon_0} \cdot \boldsymbol{v}_{\mathrm{Ph}}^2 (\boldsymbol{E} \times \boldsymbol{B})$, the vector $\boldsymbol{E}$ is perpendicular to $\boldsymbol{S}$, The dielectric displacement density $\boldsymbol{D}$ is perpendicular to $\boldsymbol{k}$, the directions of $\boldsymbol{D}$ and $\boldsymbol{E}$ include the same angle α as the directions of $\boldsymbol{k}$ and $\boldsymbol{S}$. The angle α depends on the components ε_{ik} of the dielectric tensor $\tilde{\varepsilon}$, *which in turn depend on the structure of the crystal*. If a light beam with small cross section enters a birefringent crystal of sufficient thickness the energy flux travels out of the light beam (the direction of $\boldsymbol{k}$), which implies that in this case no energy is transported.

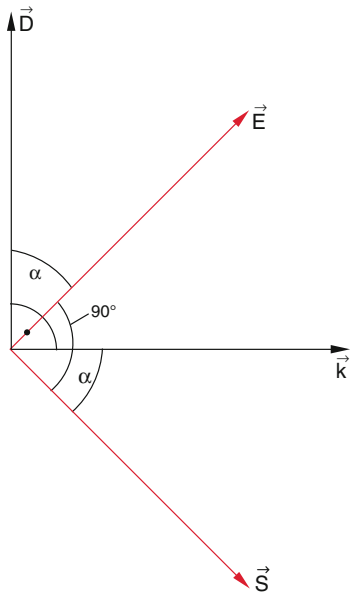


Fig. 8.38 Directions of the different vectors of a wave in anisotropic crystals. $\boldsymbol{E} \perp \boldsymbol{S}$ and $\boldsymbol{D} \perp \boldsymbol{k}$

If the optical axis coincides with the propagation direction, no birefringence occurs. Both waves have then equal refractive indices and therefore also equal phase velocities (see Fig. 8.34).

8.6 Generation and Application of Polarized Light

As has been shown in Sect. 8.7.4 an electro-magnetic wave propagating in z-direction can be always described by the representation

$$\boldsymbol{E} = \left(\boldsymbol{A}_x + \boldsymbol{A}_y\right) \mathrm{e}^{i(\omega t - kz)},$$

where the amplitudes

$$\boldsymbol{A}_x = \boldsymbol{E}_{0x} \mathrm{e}^{i\varphi_1}, \quad \boldsymbol{A}_y = \boldsymbol{E}_{0y} \mathrm{e}^{i\varphi_2}$$

are generally complex vectors.

For $\varphi_1 = \varphi_2$ the wave is *linearly polarized* (Fig. 7.4). For $|\boldsymbol{A}_x| = |\boldsymbol{A}_y|$ and $|\varphi_1 - \varphi_2| = \pi/2$ it is *circularly polarized* (Fig. 7.5) and for $|\boldsymbol{A}_x| \neq |\boldsymbol{A}_y|$ o r $|\varphi_1 - \varphi_2| \neq 0$; ½π; or π it is *elliptically polarized*.

If there is no temporary constant but a randomly fluctuating phase difference, the direction of $\boldsymbol{E}$ varies randomly in a plane perpendicular to the propagation direction z and the wave is unpolarized.

A wave, that is emitted by an oscillating dipole becomes at a sufficiently large distance r from the dipole ($r \gg d$) linearly polarized, where $\boldsymbol{E}$ is oriented parallel to the dipole axis (see Sect. 8.6.4).

Light waves are emitted by excited atoms or molecules. In most cases (for example for collisional excitation) the directions of the excited atoms are randomly distributed over all directions. Therefore the light emitted by excited atoms or molecules (e.g. in gas discharges) is unpolarized.

The question is now, how to transform this unpolarized light into polarized one. There are several experimental possibilities. Some of them will be shortly introduced in the following section [12].

8.6.1 Generation of Polarized Light by Reflection

If unpolarized light falls under the Brewster angle (Sect. 8.4.5) onto a glass plate, the reflected light contains only the component $A_{\perp}$ perpendicular to the plane of incidence (Fig. 8.14). The reflected light is therefore completely linearly polarized (see Sect. 8.4.4). The transmitted light is only partly polarized. The degree of polarization DP of partly linear polarized light is defined as

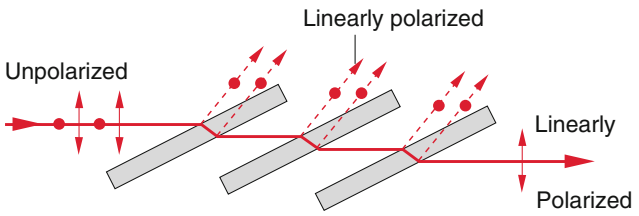


Fig. 8.39 Realization of linearly polarized light by transmission through several Brewster plates

$$\mathrm{DP} = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}}, \tag{8.80}$$

where $I_{\parallel}$ and $I_{\perp}$ are the intensities of light with the $\boldsymbol{E}$-vectors parallel rep. perpendicular to a defined direction.

From (8.65a) we can calculate the reflectivity R of the perpendicular part and with $R + T = 1$ we can conclude that the transmission T of the light at the Brewster angle is attenuated by about 15%.

The degree of polarization of the transmitted light for unpolarized incident light is at the Brewster angle

$$\mathrm{DP} = \frac{0.5 - 0.5 \cdot 0.85}{0.5 + 0.5 \cdot 0.85} \approx \sim 0.08,$$

When the incident light passes through several glass plates under the Brewster angle, the degree of polarization can be increased (Fig. 8.39). Since only the perpendicular component is reflected out of the incident beam there are no losses for the parallel component. The intensity of the transmitted light after passage through m Brewster glass plates converges towards $I_{\parallel} \rightarrow 0.5\ I_0$ for $m \rightarrow \infty$.

8.6.2 Generation of Polarized Light at the Passage Through Dichroitic Crystals

The most useful method for practical applications is the generation of polarized light by the transmission of the incident unpolarized light through thin polarization foils consisting of small dichroitic crystals, which are embedded with definite orientation in a gelatin layer. These anisotropic crystals have restoring forces for the induced atomic dipoles that depend on their direction. Therefore their resonant frequencies ω_0 in (8.21a, 8.21b) and the absorption coefficient α at a given wavelength λ depend on the direction of the electric $\boldsymbol{E}$-vector of the incident wave (Fig. 8.40). The foil can be turned in such a way, that light with the wanted polarization is transmitted and the perpendicular component is absorbed.

Such an optical anisotropy can be also realized, when a foil of cellulose hydrate is stretched into one direction, which makes it dichroitic (stress birefringence, see Sect. 8.6.6).

The drawback of the polarization foils is their relatively large attenuation even for the wanted polarization component. For high intensities (which can be achieved for instance with lasers) the large absorption leads to burning of the foil. Therefore in such cases birefringent crystals are the better choice for generating polarized light.

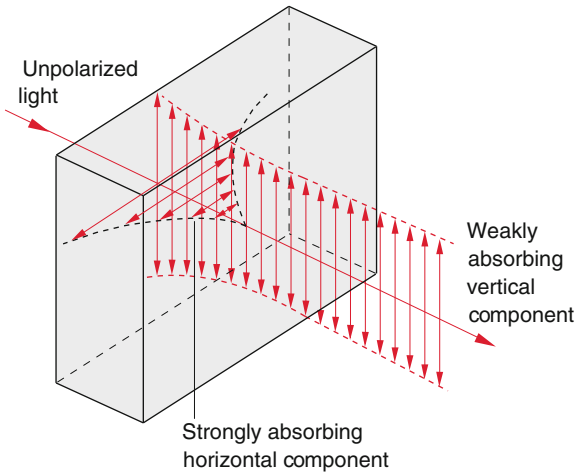


Fig. 8.40 Principle of dichroitic polarization after transmission through a dichroitic foil. One of the polarization modes is more strongly absorbed than the other

8.6.3 Birefringent Polarizers

By optical birefringence in optical uniaxial transparent crystals linearly, circular or elliptically polarized light can be generated from unpolarized light even for very high intensities.

An example is the Nicol's prism (Fig. 8.41a), which consists of a birefringent negative optical uniaxial

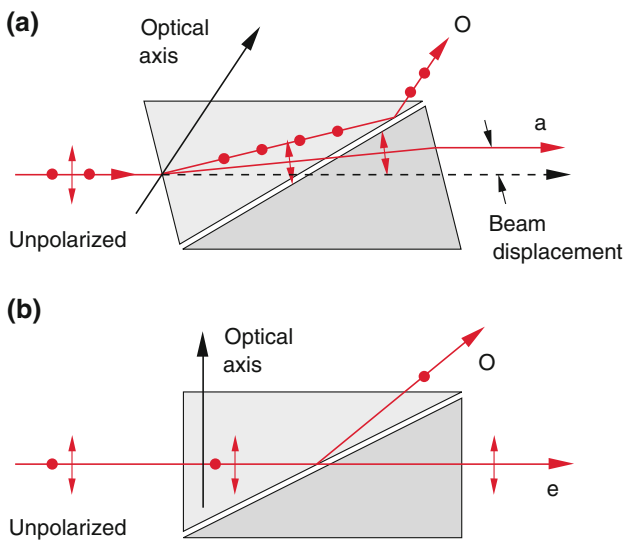


Fig. 8.41 **a**) Nicol prism for generating linearly polarized light. **b**) Glan-Thompson polarizer

rhombohedrum crystal. The crystal is cut along the diagonal surface inclined to the optical axis. The two parts are glued together by a transparent glue. If unpolarized light hits the entrance surface it is split into an ordinary and an extraordinary beam. Since $n_o > n_e$ the refracted ordinary beam suffers a larger refraction angle β. Therefore the two beams hit the glue surface under different angles. The glue (e.g. Canada balsam) has a smaller refractive index $n_C = 1.54$ than the refractive index $n_o = 1.66$ of the ordinary beam but is larger than $n_e = 1.49$ of the extraordinary beam. If the incidence angle of the ordinary beam at the interface to the Canada balsam is larger than the critical angel β_c with $\sin\beta_{c'} = n_c/n_o$ the ordinary beam is totally reflected. The transmitted light then contains only the extraordinary beam and is therefore completely linear polarized with the electric vector $\boldsymbol{E}$ parallel to the plane of incidence.

Since the entrance and exit planes of the Nicol's prism are inclined against the direction of the k-vector of the incident light a spatial displacement of the exit beam against the incident beam occurs (Fig. 8.41a).

This disadvantage is avoided for the Glan-Thompson prism (Fig. 8.41b), which has end faces perpendicular to the direction of light propagation. It is cut from a calcite crystal in such a way, that the optical axis is parallel to the end faces. Therefore there will be no birefringence for incident unpolarized light. Ordinary and extraordinary beam propagate parallel in the Glan-Thompson prism, but with different velocities $v_0 = c/n_o$ resp. $v_e = c/n_e$. At the glue layer the two beams are split, if the angle of incidence at the Canada-balsam layer exceeds the critical angle β_c of total reflection for the ordinary beam but is smaller than β_c for the extraordinary beam.

The advantages of the Glan-Thompson prism are

- there is no beam displacement as in the Nicol's prism
- the incident beam can cover the total entrance face of the prism
- the total length of the Glan-Thompson can be shorter than for the Nicol's prism.

Often it is advantageous to use both polarized beams. This can be achieved with a polarization beam splitter cube (Fig. 8.42). One could in principle use the same technique as in the Glan-Thompson prism but then a cube material has to be used which has a larger refractive index then the critical angle β_c already for $\alpha = 45°$. Therefore it is technical more efficient to use ordinary isotropic glass for the cube material and to cut the cube in the diagonal plane, place a thin polarization foil onto the cut surface and glue the two parts of the cube together. The polarizer consists of many thin dielectric layers with thickness $d = \lambda/2$, which have a high reflectivity for one polarization component but a low reflectivity for the other component (Fig. 8.43).

With birefringent crystals linear polarized light can be converted into elliptical or circular polarized light. The crystal consists of a thin plan-parallel disc with the optical axis in the plane of the disc and directed under 45° against the direction of the electric vector $\boldsymbol{E}$ of the incident light wave (Fig. 8.44).

$$\boldsymbol{E} = \boldsymbol{E}_0 \cdot e^{i(\omega t - kz)} \text{ with } \boldsymbol{E}_0 = \{E_{0x}, E_{0y}, 0\}$$

The two orthogonal components E_{0x} and E_{0y} experience different refractive indices n_1 and n_2 (see Fig. 8.34) and show after a path length d through the birefringent crystal the relative phase difference

$$\Delta\varphi = \frac{2\pi}{\lambda_0} d(n_1 - n_2).$$

When the path length d is chosen such that $\Delta\varphi = \pi/2$, the output wave is circular polarized, if the input wave is oriented under 45° against the optical axis ($E_{0x} = E_{0y}$). This polarizer is called $\lambda/4$-plate (quarter wave plate). For other angles of incidence ($\alpha \neq 45°$) is $E_{0x} \neq E_{0y}$ and the output wave is elliptical polarized.

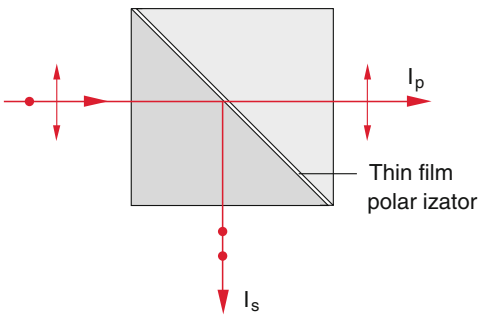


Fig. 8.42 Polarization beam splitting cube with a polarization foil inserted on the parallel and orthogonal diagonal

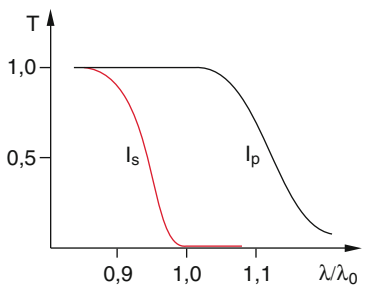


Fig. 8.43 Transmission T of the dielectric polarization beam splitter for the polarization component parallel and orthogonal to the drawing plane within a wavelength range around the optimum wavelength λ_0

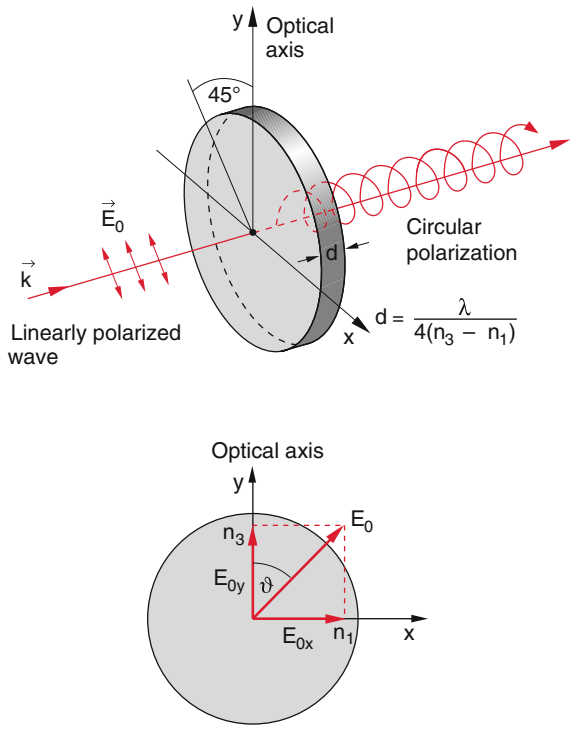


Fig. 8.44 Principle of the circular polarizer ($\lambda/4$ plate). **a**) Vivid representation **b**) direction of the electric vector $\boldsymbol{E}$ of the incident wave

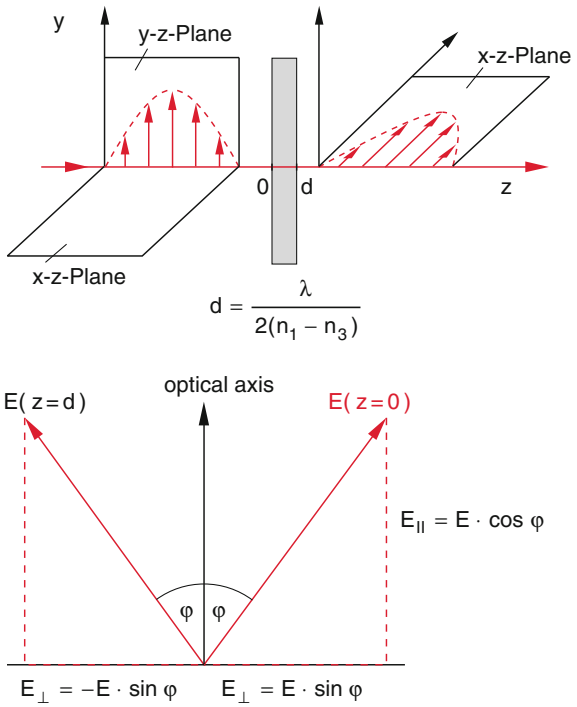


Fig. 8.45 Rotation of the polarization plane of an incident linearly polarized wave by a $\lambda/2$-plate

Example

For a positive optical uniaxial crystal with $n_1 = 1.55$ and $n_2 = 1.58$ the length of the $\lambda/4$-plate is $d = \lambda/(4 \cdot 0.03) = 8.3\ \lambda = 4.3\ \mu m$ for $\lambda = 500$ nm.

This shows that $\lambda/4$ plates are generally very thin and therefore mechanical fragile. This can be improved by choosing a crystal with a small value of Δn, which increases the thickness d. Another possibility is a thicker crystal with a higher order n of the phase difference

$$\Delta\varphi = (2m+1)\pi/2 \text{ with } m \gg 1$$

The disadvantage of these higher order $\lambda/4$-plates is the stronger dependence of the phase shift $\Delta\varphi(\lambda)$ on the wavelength λ.

8.6.4 Polarization Turners

For many applications in optics the problem arises to turn the direction of the $\boldsymbol{E}$-vector of plane polarized light by a definite angle $\Delta\alpha$. This can be achieved with a $\lambda/2$ plate which has twice the optical length n · d of a $\lambda/4$-plate. The optical axis lies in the plane of the plate. If the $\boldsymbol{E}$-vector of the incident wave forms the angle φ against the optical axis (Fig. 8.45) we can split the vector E_0 into the two components

$$E_{0\parallel} = E_0 \cdot \cos\varphi; \text{ and } E_{0\perp} = E_0 \cdot \sin\varphi$$

parallel and perpendicular to the optical axis. Both components are in phase at the entrance surface. Due to the different refractive indices the two components show at the exit surface the phase difference

$$\Delta\varphi = (2\pi/\lambda) \cdot d \cdot \Delta n.$$

For a crystal length $d = \lambda/(2 \cdot \Delta n)$ the phase difference becomes $\Delta\varphi = \pi$. With the wavenumber $k = 2\pi/\lambda$ we can write the electric vector $\boldsymbol{E}$ at the exit surface as

$$\begin{aligned} E_{\parallel} &= E_0 \cos\varphi \cdot e^{ik_{\parallel}d} e^{i\omega t}, \\ E_{\perp} &= E_0 \sin\varphi \cdot e^{ik_{\perp}d} e^{i\omega t}, \\ &= -E_0 \sin\varphi \cdot e^{ik_{\parallel}d} e^{i\omega t}, \end{aligned} \tag{8.81}$$

The vector $\boldsymbol{E}$ has turned after the crystal length d by the angle $\Delta\alpha = 2\varphi$. Turning the $\lambda/2$ plate about the direction of the incident beam allows one to realize any angle φ against the optical axis and therefore also any turning angle $\Delta\alpha = 2\varphi$.

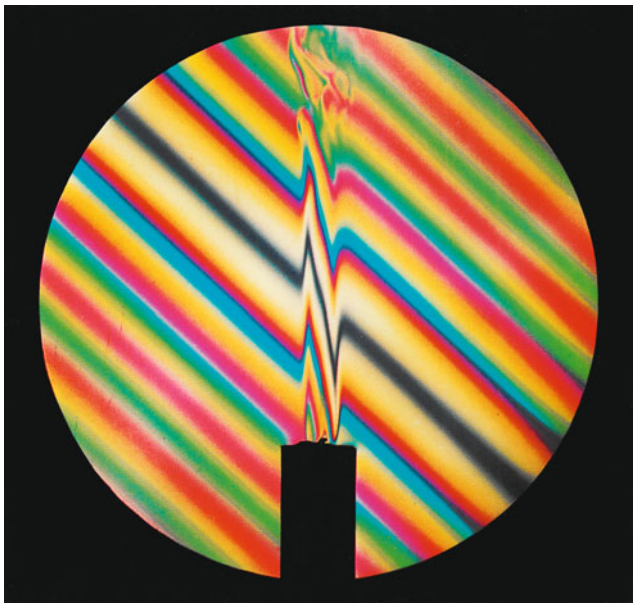


Fig. 8.46 Convection currents above a candle flame observed with a differential interferometer (interferometer with polarization) From: M. Cagnet, M. Francon, S- Malick: Atals optischer erscheinungen (Springer, Berlin, Heidelberg 1971)

The polarization characteristic of light and the influence of optical components can be used for sensitive measurements of small changes of the refractive index. One example is the polarization- dependent interferometric detection technique [13] where two applications are illustrated in Figs. 8.46 and 8.47.

Fig. 8.47 Wird 8.54x Water runner on a water surface. The interference fringes around the legs indicate the deformation of the water surface, which depend on the balance between weight of the animal and the surface tension of the water. From Francon and Malik (1971)

8.6.5 Optical Activity

Some materials turn the polarization plane even for an arbitrary direction of the ***E***-vector after passing through the material thickness d by the angle

$$\alpha = \alpha_s \cdot d. \tag{8.82}$$

The proportionality factor α_s is the **specific optical rotation power** (Fig. 8.48). One has to distinguish between right and left turning materials. The sense of rotation is defined for an observer looking towards the propagation direction of light. They are labeled as "+" for right turning materials and "–" for left turning ones.

The physical reason for the optical activity are symmetry properties of the optical medium. For some substances optical activity is observed only in the solid crystalline phase whereas it disappears in the liquid or gaseous phase. It therefore must be caused by the special symmetry of the crystal. One example is crystalline quartz, which can be found in nature as right- or left turning quartz (Fig. 8.49).

On the other side there are also substances (e.g. sugar or lactic acid) which show optical activity also in the liquid

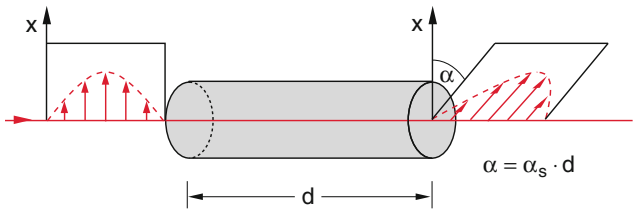


Fig. 8.48 Optical activity of a medium, indicating the rotation of the polarization plane

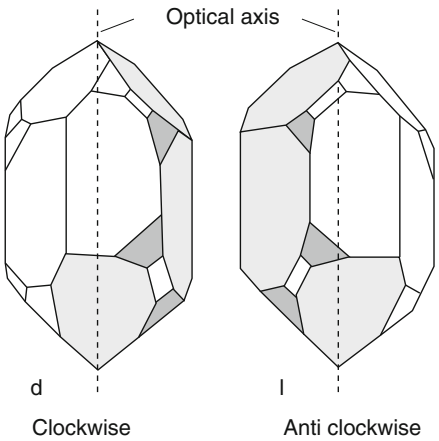


Fig. 8.49 The two mirror images of the crystal structures of a left- and right turning quartz crystal

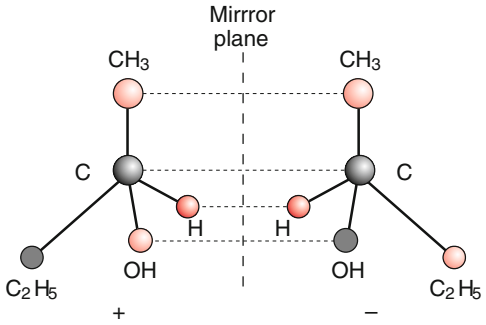


Fig. 8.50 Two isomeric forms of the 2-butanol molecule which are mirror images with respect to a mirror plane perpendicular to the drawing plane

phase. Here the symmetry of the molecules must be responsible for the optical activity (Fig. 8.50).

A complete explanation of optical activity is only possible on the basis of quantum theory. However, a descriptive model can illustrate the physical basis, as is shortly outlined as follows [14]:

Similar to the generation of linearly polarized waves by the induced oscillations of atomic dipoles in a homogeneous medium, here it is assumed that the outer electrons of these special molecules or crystals are induced by a circular polarized wave to elliptical motions about the propagation direction of the wave. This model is supported by the spiral shaped arrangement of oxygen and carbon atoms in crystalline quartz. The spiral is right-handed for right turning quartz and left-handed for left-turning quartz. Such molecules are called **chiral molecules**. They exist in two mirror configurations (**mirror isomers**, Fig. 8.50). Examples are sugar, lactic acid or 2-butanol.

We can compose a linear polarized wave

$$\boldsymbol{E} = \hat{\boldsymbol{e}}_x E_{0x} \cdot \mathrm{e}^{\mathrm{i}(\omega t - kz)}$$

of two opposite circular polarized waves

$$\begin{aligned} \boldsymbol{E}^+ &= \frac{1}{2}\left(\hat{\boldsymbol{e}}_x E_{0x} + i\hat{\boldsymbol{e}}_y E_{0y}\right)\mathrm{e}^{\mathrm{i}(\omega t - kz)} r \\ \boldsymbol{E}^- &= \frac{1}{2}\left(\hat{\boldsymbol{e}}_x E_{0x} - i\hat{\boldsymbol{e}}_y E_{0y}\right)\mathrm{e}^{\mathrm{i}(\omega t - kz)}. \end{aligned} \tag{8.83}$$

If the two circular polarized components have different phase velocities $v^+ = c/n^+$ or $v^- = c/n^-$ the composite wave becomes again linear polarized after the pass length $d = \lambda/\Delta n$ but its plane of polarization has turned by the angle

$$\alpha = \frac{\pi}{\lambda_0} d(n^- - n^+)$$

The differing refractive indices n^+ and n^- are due to the different interaction of the right- or left circular polarized wave with the electrons that move in a preferential direction of rotation.

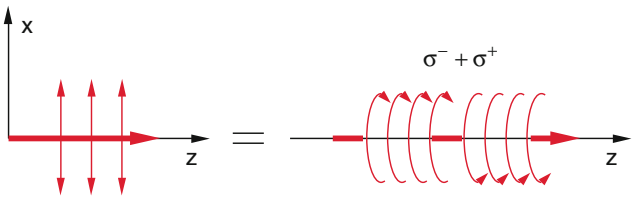


Fig. 8.51 A linearly polarized wave can be composed of a right (σ^-) and a left (σ^+) circular polarized wave

The fact that optical activity can also occur in a liquid where the orientations of the molecules are randomly distributed without the incident optical wave, can be understood as follows:

Due to the electric and magnetic dipoles of chiral molecules, induced by the circular polarized wave a small orientation is generated which can effect optical activity. For biological molecules nature apparently favors one of the two mirror isomers.

For instance, the blood sugar is always left handed. With a polarimeter the rotation angle $\alpha = \alpha_s \cdot C \cdot d$ of a sugar solution with the concentration C and the length d can be measured and the concentration C can be determined. The sample is placed between two crossed polarizers (Fig. 8.52) and the analyzer is turned by the angle $-\alpha$ until the transmitted intensity becomes again zero [13].

8.6.6 Stress Birefringence

Even in homogeneous isotropic media optical birefringence can be induced by anisotropic external pressure or tension, which result in changes Δn of the refractive index n which are dependent on orientation and location in the media. Measuring these changes Δn gives information about the spatial distribution of mechanical stress in the medium. Such a measurement can be performed with the design shown in Fig. 8.52 where the white light beam is expanded to a cross section that covers the whole sample. The transparent sample is placed between to crossed polarizers. For an isotropic sample no light is transmitted through the second crossed polarizer and the observation plane is dark. When mechanical stress is applied to the sample a spatially dependent change of the polarization direction is induced and colored pattern is observed behind the second polarizer which images the spatial distribution of the anisotropic stress (Fig. 8.53). The optical phase shift.

$$\Delta\varphi(x,y) = \frac{2\pi}{\lambda_0} \cdot \int_0^d \Delta n(x,y)\,dz$$

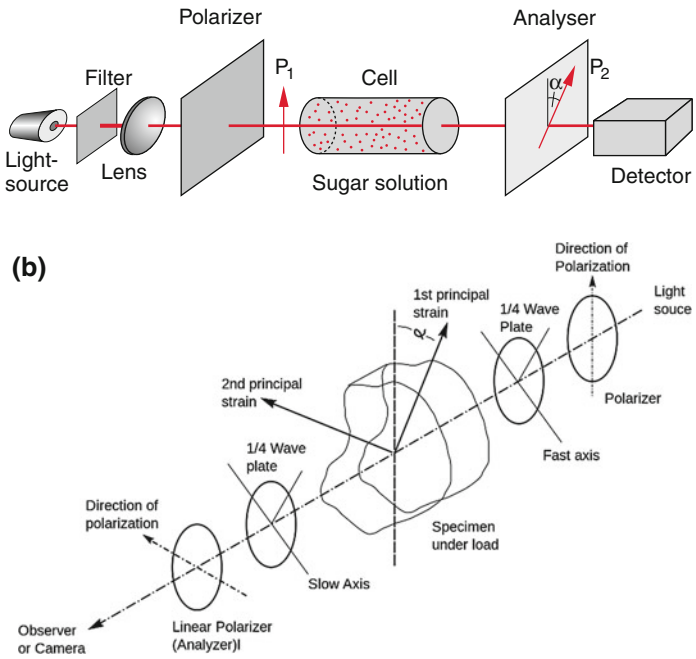


Fig. 8.52 **a**) Measurement of the sugar concentration with a polarimeter. **b**) Measurement of stress-induced birefringence

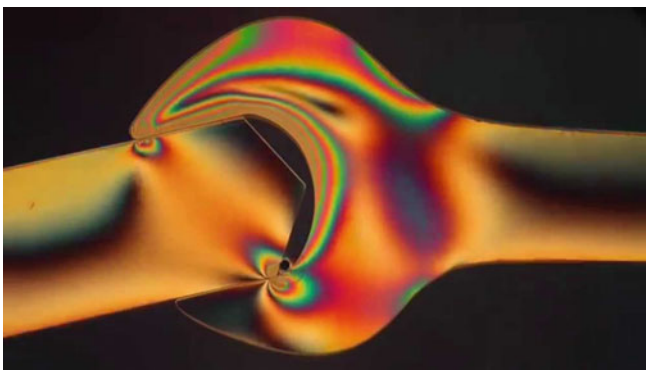


Fig. 8.53 Stress induced birefringence of a spanner holding a work piece (Dr. G. Haberland, Woltersdorf)

is given by the integral of Δn integrated over the path length $z = d$ through the sample. Since $\Delta\varphi$ depends on the wavelength, measurements with white light give colored patterns of the spatial distribution of stress-induced birefringence and allows a detailed information about the distribution and magnitude of the stress. Glass blowers use this technique of polarimetry to judge the residual stress in blown glass, which can be removed by annealing the glass, where the temperature is raised up to a value closely below the melting temperature and then slowly cooled down.

Illustrative examples are a plexi glass rod which is supported by two holders at the two ends and pressed down in the middle. Another nice example is the water strider on the water surface shown in Fig. 8.47, which does not sink due to surface tension, but causes dents of the water surface around the legs caused by the weight of the strider.

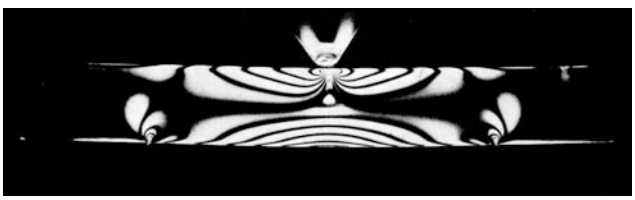


Fig. 8.54 Birefringence of a plexi glass rod induced by mechanical stress in the rod supported at two points and pressed in the midpoint

For many technical applications the spatial stress distribution is very important to judge the upper load limit for a bridge or a building. For the measurements a transparent reduced scale model of the building in question is placed

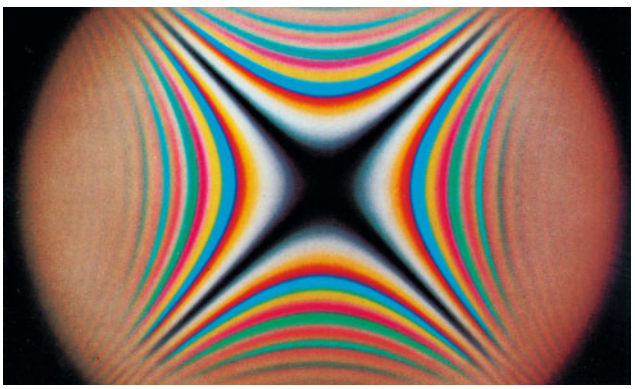


Fig. 8.55 Stress-induced birefringence detected in convergent light. Two equal quartz plates cut parallel to the optical axis are turned to a crossed position of their optical axes and are placed between two crossed polarizers. Convergent white light passes through the arrangement. From Francon et al. (1971)

between the crossed polarizers. In Fig. 8.52b the arrangement for measuring stress-induced birefringence is shown and Fig. 8.54 illustrates the observed pattern for a plexi glass rod under specific stress.

In Fig. 8.53 the stress-induced polarization of a spanner is shown which illustrates the spatial distribution of the stress and in Fig. 8.55 the stress-induced birefringence in a quartz plate is detected between two crossed polarizers.

8.7 Nonlinear Optics

For sufficiently small electric field strength $\boldsymbol{E}$ of the incident wave the amplitudes of the induced oscillations of the atomic electrons are small and the restoring forces are proportional to the elongations from the equilibrium position (Hooke's linear range). The induced electric dipole moments $\boldsymbol{p} = \alpha \cdot \boldsymbol{E}$ are proportional to the electric field $\boldsymbol{E}$. The components of the dielectric polarization

$$P_i = \varepsilon_0 \cdot \sum_j \chi_{ij} E_j$$

depend linearly on $\boldsymbol{E}$. The coefficients χ_{ij} are the components of the electric susceptibility tensor $\tilde{\chi}$ (see 1.58). This is the range of **linear optics**. For isotropic media the susceptibility tensor reduces to a number i.e. $\chi_{ij} = \chi \cdot \delta_{ij}$ where χ is a scalar quantity.

Example

The electric field strength $\boldsymbol{E}$ of the sunlight within a spectral range $\Delta\lambda = 1$ nm around $\lambda = 500$ nm reaching the earth surface is about 3 V/m. The Coulomb force which binds the outer atomic electron to the nucleus is, however,

$$E_C \approx \frac{10\,\mathrm{V}}{10^{-10}\,\mathrm{m}} = 10^{11}\,\mathrm{V/m}. \tag{8.84}$$

Therefore the elongation of the atomic electrons induced by the sunlight, are very small compared to their average distance from the atomic nucleus and the range of linear optics is not exceeded.

For much larger intensities [as for example reached with focused laser beams (see Vol. 3)] the range of nonlinear elongations of the atomic electrons can be reached. Instead of (8.84) we have to use the equation

$$P_i = \varepsilon_0 \left(\sum_j \chi_{ij}^{(1)} E_j + \sum_j \sum_k \chi_{ij}^{(2)} E_j E_k + \sum_j \sum_k \sum_l \chi_{ijkl}^{(3)} E_j E_k E_l + \ldots \right), \tag{8.85}$$

where $\chi^{(n)}$ is the susceptibility of n-th order, which is described by a tensor of rank $(n + 1)$. Although the quantities $\chi^{(n)}$ which depend on the symmetry properties of the medium rapidly decrease with increasing n the higher order terms in (8.85) can no longer be neglected for high intensities of the incident wave.

When a monochromatic light wave

$$\boldsymbol{E} = \boldsymbol{E}_0 \cdot \cos(\omega t - kz) \tag{8.86}$$

propagates through the medium, the polarization P contains, due to the higher order powers E^m of the electric field E with $m > 1$, besides the fundamental frequency ω also higher order frequencies $m\omega$ ($m = 2; 3; 4; \ldots$). This implies, that the induced oscillating dipoles emit electromagnetic waves not only on the fundamental frequency ω (Rayleigh scattering) but also on higher harmonics $m\omega$. The amplitudes of these higher harmonics depend on the coefficients $\chi^{(m)}$ (i.e. from the characteristic features of the medium) but also on the amplitude of the incident wave.

We will illustrate this by some examples [15, 16].

8.7.1 Optical Frequency Doubling

Inserting (8.86) into (8.85) gives when neglecting all terms $\chi^{(m)}$ with $m > 2$ The polarization at $z = 0$

$$P_x = \varepsilon_0 \left(\chi^{(1)} E_{0x} \cos\, \omega t + \chi^{(2)} E_{0x}^2 \cos^2\, \omega t \right).$$

Here we have assumed an isotropic medium and a linear polarized incident wave with $\boldsymbol{E_0} = \{E_{0x}, 0, 0\}$.

With $\cos^2 x = ½(1 + \cos 2x)$ we obtain

$$P_x = \varepsilon_0 \left(\chi^{(1)} E_{0x} \cos\, \omega t + \frac{1}{2}\chi^{(2)} E_{0x}^2 + \frac{1}{2}\chi^{(2)} \cos^2\, \omega t \right). \tag{8.87}$$

The polarization contains a constant term $½\varepsilon_0 \chi^{(2)} E_{0x}^2$, a term that oscillates with the frequency ω and a term that describes the oscillation with the double frequency 2ω. This means:

Every atom hit by the incident wave with the fundamental frequency ω, emits a scattered wave with the frequency ω (Rayleigh scattering) and a part that oscillates with the frequency 2ω (overtone wave).

The amplitude of this overtone wave is, according to (8.87) proportional to the square of the amplitude of the incident wave, i.e. the intensity of the overtone wave is proportional to the square of the incident intensity.

The microscopic parts from the different atoms superimpose each other. This superposition can only lead to a macroscopic wave if all the different parts are in phase at each location in the nonlinear medium. This demands that the phase velocities of the fundamental wave and the overtone wave must be equal. Because of dispersion this cannot be achieved in a normal isotropic medium, where the phase velocity depends on the frequency of the wave, but can be realized in birefringent optical crystal where the phase velocity depends on the propagation direction against the optical axis and on the polarization of the wave.

8.7.2 Phase Matching

When the plane wave (8.86) propagates in z-direction through the medium, it induces in each plane $z = z_0$ atomic dipoles. The oscillation phase of the dipoles depends on the phase of the inducing wave at the plane $z = z_0$. In the neighboring plane $z = z_0 + \Delta z$ the same phase difference between inducing wave and dipole oscillation exists.

The secondary wave emitted by the dipoles at the frequency ω reaches the plane $z = z_0 + \Delta z$ at the same time as the inducing wave. They therefore superimpose the secondary waves generated in the plane $z = z_0 + \Delta z$ in phase (see Sect. 8.1). This leads to a macroscopic secondary wave at the frequency ω that superimposes the primary wave at ω. Because of its phase shift against the primary wave the total wave at the frequency ω has a phase velocity $v_{ph} - c/n$ that is smaller than in vacuum (see Sect. 8.1).

Due to the dispersion of the medium this is generally no longer true for the overtone waves, because the phase velocities $v_{ph}(2\omega) = c/n(2\omega) \neq v_{ph}(\omega)$ are different for the fundamental and the overtone wave. Therefore the overtone wave generated at the plane $z = z_0$ arrives at the plane $z_0 + \Delta z$ later than the fundamental wave and the overtone wave at 2ω, generated at $z = z_0$ has a phase lag against the overtone wave generated in $z = z_0 + \Delta z$ when it arrives at $z = z_0 + \Delta z$. This means: In isotropic media the microscopic shares of the overtone wave generated at the different atoms cannot build up to a macroscopic overtone wave (Fig. 8.56).

After a path length

$$\Delta z = \frac{\lambda/2}{n(2\omega) - (n\omega)} = l_c$$

the overtone wave shows a phase lag of $\Delta\varphi = \pi$. It is therefore opposite in phase against the secondary waves emitted at the plane $z = l_c$ and the superposition becomes destructive at the critical phase-matching length l_c.

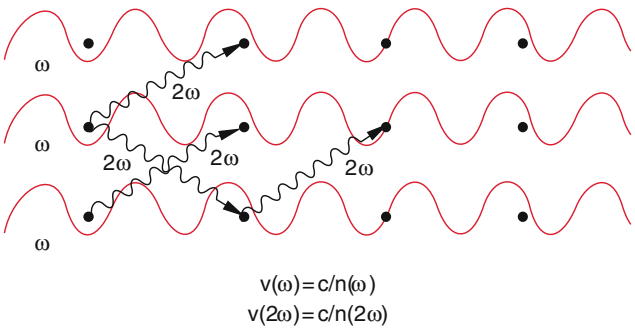


Fig. 8.56 Schematic representation of the realization of optical frequency doubling

Averaged over the whole medium the conversion of the fundamental into the overtone wave becomes zero, i.e. there is practical no energy transfer from the incident fundamental wave into the overtone wave. One therefore has to look for a way where the phase velocities of fundamental and overtone wave become equal.

Fortunately birefringent crystals offer such a possibility (see Sect. 8.5), based on the different phase velocities of ordinary and extraordinary wave. If it is possible to choose a direction θ_p against the optical axis of an uniaxial crystal where the refractive index $n_e(2\omega)$ of the extraordinary wave is equal to $n_0(\omega)$ of the ordinary wave the fundamental wave $\boldsymbol{E_0}(\omega)$ propagates in this direction with the same velocity as the overtone wave $\boldsymbol{E}_e(2\omega)$ (Fig. 8.57). Now all overtone waves generated in arbitrary planes can superimpose in phase with the overtone waves in subsequent planes. In this case a macroscopic overtone wave is generated which propagates into the

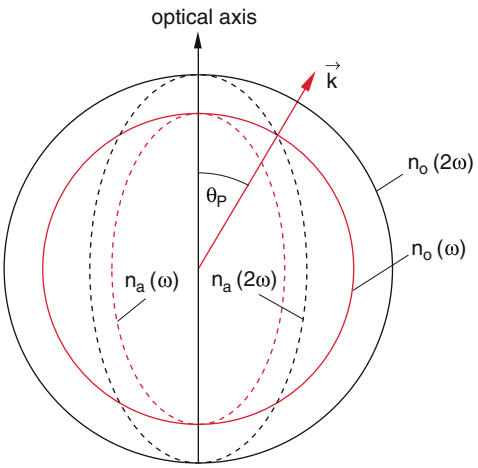


Fig. 8.57 Phase-matching of fundamental wave with frequency ω and its first harmonic (2ω) in birefringent crystals

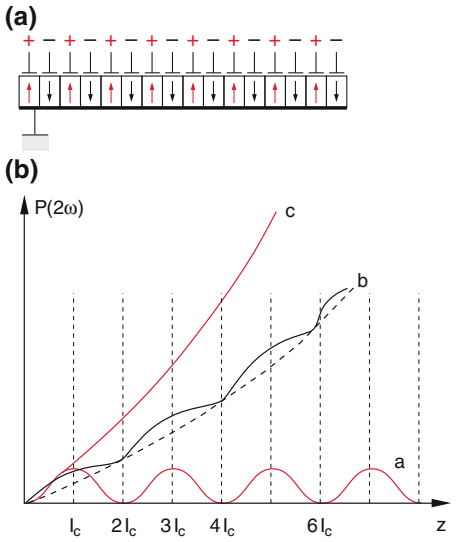


Fig. 8.58 Quasi-phase-matching in periodical poled optical crystals. **a**) periodical change of the difference $\Delta n = n(2\omega) - n(\omega)$ in a ferro-electric crystal, **b**) Output power $P(2\omega)$ as a function of the crystal length. *a* single crystal with slight phase- mismatching *b* periodical poled crystal with the same phase mismatch *c* single crystal with ideal phase matching (is only valid for the correct frequency ω)

same direction as the fundamental wave (**optical frequency doubling**). For example is the red light at $\lambda = 690$ nm emitted by the Ruby Laser converted into ultraviolet light at $\lambda = 345$ nm. Since the ordinary and the extraordinary wave are polarized perpendicular to each other is

$$\boldsymbol{E}(2\omega) \perp \boldsymbol{E}(\omega).$$

The phase-matching condition is

$$\begin{aligned} n_a(2\omega) = n_0(\omega) &\Rightarrow v_{ph}(\omega) = v_{ph}(2\omega) \\ &\Rightarrow \boldsymbol{k}(2\omega) = 2\boldsymbol{k}(\omega). \end{aligned} \tag{8.88}$$

The disadvantage of birefringent crystals for optical frequency doubling is the limited spectral range of the phase-matching condition for a given angle θ_p against the optical axis, which is only strictly fulfilled for a selected wavelength. For other wavelengths the optical axis has to be turned.

In recent years another method of **quasi-phase-matching** has been introduced. Here a ferro-electric medium is used which consists of many thin slices with alternatively changing signs of the refractive index difference $\Delta n = n(2\omega) - n(\omega)$. This alternation is realized by an external electric field which periodically changes its polarity (Fig. 8.58a). The phase difference developing within one slice is compensated in the next slice. The intensity of the overtone wave does not increase as fast with the length of the doubling device as for exact phase matching in a single crystal, but quasi phase matching can be realized for a much larger wavelength range [17].

8.7.3 Optical Frequency Mixing

When two light waves

$$\begin{aligned} \boldsymbol{E}_1 &= E_{01}\hat{\boldsymbol{e}}_x \cos(\omega_1 t - \boldsymbol{k}_1 \cdot \boldsymbol{r}) \\ \boldsymbol{E}_2 &= E_{02}\hat{\boldsymbol{e}}_x \cos(\omega_2 t - \boldsymbol{k}_2 \cdot \boldsymbol{r}) \end{aligned}$$

are superimposed in a nonlinear optical medium, the total electric field $\boldsymbol{E} = \boldsymbol{E_1} + \boldsymbol{E_2}$ causes, according to (8.85) a dielectric polarization P of the atoms, with a nonlinear part

$$\begin{aligned} P^{(2)}(\omega) &= \varepsilon_0 \chi^{(2)} \left[E_{01}^2 \cos^2 \omega_1 t + E_{02}^2 \cos^2 \omega_2 t + 2E_{01}E_{02} \cos \omega_1 t \cdot \cos \omega_2 t\right] \\ &= \frac{1}{2}\varepsilon_0 \chi^{(2)} \left[\left(E_{01}^2 + E_{02}^2\right) + E_{01}^2 \cos 2\omega_1 t + E_{02}^2 \cos 2\omega_2 t \right. \\ &\quad \left. + 2E_{01}E_{02}(\cos(\omega_1 + \omega_2)t + \cos(\omega_1 - \omega_2)t)\right]. \end{aligned} \tag{8.89}$$

Besides the overtones with $\omega = 2\omega_1$ resp. $2\omega_2$ waves with the sum frequency $(\omega_1 + \omega_2)$ and the difference frequency $(\omega_1 - \omega_2)$ are generated.

Choosing the right phase matching one can realize that for one of these frequencies all contributions of the secondary waves from the different atoms in the medium superimpose with the correct phase, resulting in a macroscopic wave with the corresponding frequency (**optical frequency mixing**).

For instance is the phase matching condition for the generation of the sum frequency

$$\begin{aligned} &\boldsymbol{k}_3(\omega_1 + \omega_2) = \boldsymbol{k}_1(\omega_1) + \boldsymbol{k}_2\omega_2 \\ &n_3 \cdot \omega_3 = n_1\omega_1 + n_2\omega_2 \quad \text{with} \quad n_i = n(\omega_i). \end{aligned} \tag{8.90}$$

It is generally easier to accomplish phase-matching for the sum frequency generation than for optical frequency doubling, because the directions of the wave vectors $\boldsymbol{k}_1$ and $\boldsymbol{k}_2$ can be chosen within certain limits in order to achieve phase matching for a wider frequency range.

This optical frequency mixing has considerably enlarged the spectral range for the realization of coherent radiation sources, extending from the near infrared to the ultraviolet region. Furthermore it has allowed the investigation of electronic properties of nonlinear optical media [18, 19].

Since the efficiency of these nonlinear optical frequency mixing increases with the square of the incident intensity of the fundamental wave, such mixing experiments were in the earlier days of lasers only possible with pulsed lasers with high peak intensities (see Vol. 3 and [20]). Meanwhile new optical nonlinear crystals can be grown (e.g. barium-beta Borat $Ba(BO_2)_2$ or lithium-iodate $LiJO_3$) with large nonlinear coefficients of the susceptibility tensor $\chi^{(2)}$ which enable optical frequency mixing or doubling even with continuous (cw) lasers.

8.7.4 Generation of Higher Harmonics

In recent years researchers have succeeded in the generation of high harmonic frequencies $m\omega$ ($m = 2; 3; 4; \ldots 300$) by focusing pulsed lasers into a container with a noble gas (e.g. neon or argon) at high pressures. The atoms in the gas become ionized in the high electric field of the focused laser beam. The photo- electrons are periodically accelerated by the electric field $E(\omega)$ of the laser, which changes sign with the optical frequency ω. These accelerated electrons radiate electromagnetic waves within a broad frequency range [21]. In Fig. 8.59 such a frequency spectrum of higher harmonics is shown [18] generated by a focused fundamental laser wave at $\lambda = 1.050$ nm. The energy of 160 eV corresponds to the 150th overtone wave at $\lambda = 7$ nm which lies already in the soft X-ray region.

Example

For $m = 100$ the frequency $m\omega$ of the overtone wave from the fundamental wave at $\lambda = 700$ nm ($\omega = 2\pi \times 10^{14}$ s^{-1}) is $2\pi \times 10^{16}$ s^{-1}, corresponding to a wavelength $\lambda_m = 7$ nm which is located in the soft X-ray region. This illustrates that the high harmonics generation enables the realization of intense X-ray sources.

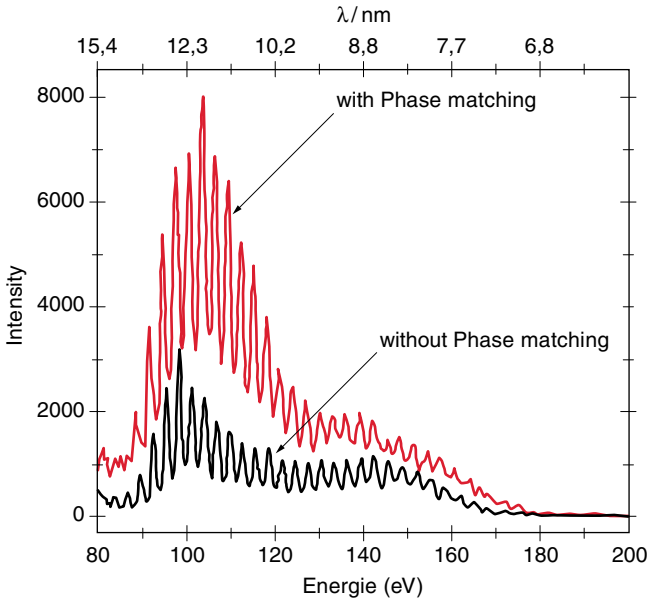


Fig. 8.59 Spectrum of high harmonics of the laser radiation at $L = 1.05$ µm with quasi. phase-matching (red curve) and without phase matching (black curve)

Summary

- Electromagnetic waves have in a medium with refractive index n the phase velocity $v_{ph} = c/n$, which depends on the frequency ω, because $n(\omega)$ depends on ω.
- The refractive index is a complex number

$$n = n_r - i\kappa.$$

The real part describes the dispersion, the imaginary part the absorption of an incident electromagnetic wave. The two parts n_r and κ are related by the dispersion relations (8.21a).

- The intensity of an electro-magnetic wave propagating into the z-direction through a medium with absorption coefficient α decreases with z as

$$I = I_0 \cdot e^{-\alpha z} \text{ with } \alpha = (4\pi/\lambda_0)\kappa.$$

This Beer's absorption law is valid for not too high intensities, where saturation effects can be still neglected.

- At the interface between two media with different refractive indices n_1 and n_2 reflection and refraction occur. Amplitudes and polarization of reflected and refracted waves depend on the angle of incidence α and on the polarization of the incident wave. They can be calculated using the Fresnel formulas (8.61a, 8.61b, 8.62a, 8.62b).
- The sum of reflectivity R and transmission T in absorption-free media is always

$$R + T = 1$$

For vertical incidence ($\alpha = 0°$) is

$$R = \left|\frac{n_1 - n_2}{n_1 + n_2}\right|^2.$$

Interfaces of strongly absorbing media have a high reflectivity.

- For the transition from an optically dense to an optically thin medium total reflection occurs for angles $\alpha > \alpha_c$. Nevertheless the wave penetrates into a thin layer $\Delta x < \lambda$ of the optically thinner medium as evanescent wave. The total reflection is still 100%.
- At the Brewster angle $\alpha = \alpha_B$ the reflectivity $R_\parallel$ for the component $A_\parallel$ parallel to the incidence plane becomes zero. For $\alpha = \alpha_B$ the direction of the reflected beam is perpendicular to that of the refracted beam.
- In anisotropic media the electric field vector $\boldsymbol{E}$ and the dielectric displacement vector $\boldsymbol{D}$ are generally no longer parallel. The pointing vector S includes with the wave vector $\boldsymbol{k}$ the same angle α as $\boldsymbol{E}$ with $\boldsymbol{D}$.

In birefringent media the incident wave splits into an ordinary and an extra-ordinary part. The refractive index depends on the polarization of the incident wave. For the ordinary wave n_o is independent of the direction of $\boldsymbol{k}$, like in isotropic media. For the extraordinary wave n_e depends on the angle between $\boldsymbol{k}$ and the optical axis.

- Polarized light can be generated

  (a) by reflection under the Brewster angle
  (b) by dichroitic thin film polarizers
  (c) by birefringent crystals.

- Electromagnetic waves in media can be described by a wave equation that is derived from the Maxwell equations. The equations contain for the propagation in media an additional term that represents the polarization of the medium by the wave. This polarization is the source of new waves (secondary waves), emitted by the induced atomic dipoles.
- If the incident light has sufficiently high intensity, the linear dependence of the oscillation amplitude of the dipoles for small elongations is exceeded and nonlinear effects arise. The dipoles emit overtone waves with frequencies mω, (m = 1; 2; 3; …). For the correct alignment of the nonlinear optical birefringent crystal the secondary overtone waves superimpose in phase (phase matching) and a macroscopic overtone wave is generated (optical frequency doubling, nonlinear optics).

Problems

8.1 Calculate the refractive index of air at atmospheric pressure for light with the wavelength $\lambda = 500$ nm, using Eq. (8.12b). The resonance frequency of the nitrogen molecules is $\omega_0 = 10^{16}$ s^{-1}. The influence of the other gases should be neglected. Comparing with the value in Table 8.1 what can you say about the oscillator strength in Eq. (8.13)?

8.2 Under which angle must a light beam enter the interface air-glass that the angle between incident and reflected beam becomes equal to the angle between incident and refracted beam?

8.3 Assume that 8 atoms are located at the 8 corners of a cube with side length $L = 100$ nm. An incident plane light wave propagating into the z-direction induces the atoms to oscillations in x-direction. How large is the fraction of the incident light which is scattered into the y-direction when the scattering cross section for a single atom is $\sigma = 10^{-30}$ m^2?

8.4 Derive the Fresnel Eqs. (8.62a, 8.62b)

8.5 Calculate the amplitude reflection coefficients $_{\parallel}$ and $_{\perp}$ and the reflectivity R at the interface between air ($n_1' = 1$, $\kappa = 0$) and silver ($n_2' = 0.17$, $\kappa = 2.94$) for the angles of incidence $\alpha = 0°$, 45° and 85°, using the Fresnel formulas (8.62a, 8.62b).

8.6 A light wave with the power $P = 1$ W passes through an absorbing medium with the length $L = 3$ cm and the absorption coefficient α. How large is the absorbed power for

(a) $\alpha = 10^{-3}$ cm^{-1}
(b) for $\alpha = 1$ cm^{-1}?

8.7 An optical fiber has the kernel diameter of 10 μm. The refractive index of the kernel is $n_1 = 1.6$, that of the cladding $n_2 = 1.59$. What is the minimum radius of curvature of the fiber in order to maintain total reflection?

8.8 Show that Eq. (8.12a) could be also written for $\omega - \omega_0 \gg \gamma$ as $n - 1 = a + b/(\lambda^2 - \lambda_0^2)$ in order to obtain a simple dispersion formula for air at atmospheric pressure.

8.9 An optical wave with the frequency $\omega = 3.5 \times 10^{15}$ s^{-1} ($\lambda = 500$ nm) and the intensity $I = 10^{12}$ W/m^2 travels through a nonlinear uniaxial crystal with the nonlinear susceptibility $\chi^{(2)}(\omega) = 8 \times 10^{-13}$ m/V. the refractive indices are $n_0(\omega) = 1.675$; $n_e(2\omega, \theta = 90°) = 1.615$; $n_0(2\omega) = 1.757$

(a) For which angle θ_{opt} against the optical axis can be phase matching obtained?
(b) What is the coherence length $L_{coherent}$ for a small misalignment $\theta = \theta_{opt} + 1°$?
(c) What is the output intensity I(2ω), which is given by the relation

$$I(2\omega, L) = I^2(\omega) \cdot \frac{2\omega^2 \left|\chi^{(2)}\right|^2 L^2}{n^3 c^3 \cdot \varepsilon_0} \frac{\sin^2(\Delta k \cdot L)}{(\Delta k \cdot L)^2}$$

for $L = L_{coherent}$?

References

1. J.W. Robinson: Atomic Spectroscopy (Marcel Dekker, 2013)
2. J.D Jackson: Classical Electrodynamics 3rd ed. (Wiley 1998)
3. C.S-.Bloch: Eight Velocities of light. Am. J. Physics **45**, 538 (1977)
4. R.L. Smith: The velocities of Light: Am. J. Physics **38**, 978 (1970)
5. L. Hau et.al.:Light speed reduction to 17 m/s in an ultracold atomic gas. Nature **397**, 594 (1999)
6. G. Dolling, M. Wegener, S. Linden: Negative index metamaterial at 780 nm. (Opt.Express **15**, 11536 (2007))
7. J.B-. Pendry: negative refraction makes a perfect lens. Phys. Rev. Letters **85**, 3966 (2000)
8. S. Anantha Ramaskrishna: Physics of negative refractive index Reports on Progress in Physics **68** (2), 4549–521, (2005)
9. M.V. Klein, Th.E. Furtak: Optics (Wiley 1986)
10. John D. Joannopoulos, Steven G. Johnson Joshua N. Winn Robert D. Meade: Photonic Crystals (Princeton Univ. Press, 2nd ed. 2008)
11. D.W. Prather et.al, Photonic crystals Theory, Applications and Fabrication (Wiley 2009)
12. M.W. McCall , Ian J. Hodgkison, Qihong Wu: Birefringent thin films and polarizing elements 2nd ed. (World Scientific Singapore 2015))
13. M. Francon, S. Mallik: Polarization Interferometers (Wiley, London 1961)
14. St.F. Mason: Molecular Optical Activity and the Chiral Discrimination (Cambridge Univ. Press 1982)
15. G.C. Baldwin: An Introduction to Nonlinear Optics (Plenum Press new York 1969 and Springer Heidelberg 2013)
16. R.W. Boyd: Nonlinear Optics 2nd ed. (Academic Press 2002)
17. M.M. Feyer et.al. Quasi-phase-matchned second harfmonic generation. IEEE J. Quant. Elctr. **QE 28**, 2631 (1992)
18. J.F. Ward: Resonant and non-resonant optical frequency mixing in simple molecular systems. (Springer Series in Optical Sciences 164, (Springer 2016)
19. A. Smith, Crystal Nonlinear Optics with SNLO examples (AS Photonics 2016)
20. Y.R. Shen: The principles of Nonlinear Optics (Wiley Series in Pure and Applied Optics 1984)
21. F.X. Kaertner: High Harmonics Generation driven by a visible non-collinear optical parametric amplifier (LAP LAMBERT Academic Publishing 2010)

Geometrical Optics

9

For many applications the wave nature of light is of minor importance. The main interest is the propagation direction of light and its alterations by imaging elements, such as mirrors, prisms and lenses.

The propagation direction of a wave is determined by the normal vector to the phase front. These normal vectors $\boldsymbol{k}(\boldsymbol{r})$ as a function of the location along the propagation of the light wave form the **light rays** in geometrical optics.

When a light wave is limited by boundaries, such as apertures, edges of lenses or mirrors we call the confined part of the wave a **light beam**. A light beam can be regarded as the total quantity of all light rays filling the cross section of the light beam (Fig. 9.1). Besides its cross section and its propagation direction we can attribute to the light beam also wave qualities, such as wavelength λ, amplitude E, velocity $c' = c/n$, intensity $I = c' \cdot \varepsilon \cdot \varepsilon_0 \cdot E^2$ and polarization. In a more sloppy language one speaks of an intense or a weak light beam or a polarized light ray.

The description of a light wave confined by limiting boundaries by rays of light is of course an approximation. Inside the light beam, where the change of the electric field E across the beam is sufficiently slow and therefore negligible, this approximation is justified (For a plane wave $\boldsymbol{E} = \boldsymbol{E_0} \cdot \cos(\omega t - kz)$ $\boldsymbol{E}$ is constant on a plane z = constant, i.e. along the x- and y-directions). However, at the edges of the light beam large changes of the intensity appear and diffraction effects are no longer negligible.

We can use the approximation of geometrical optics, if the diameter of the light beam is large compared to the wavelength λ. In this case diffraction effects can be neglected.

As a rule of thumb note that for λ = 500 nm light beams should have a diameter $d > 10\,\mu$m in order to treat them by geometrical optics, because then diffraction effects can be neglected. In this sense light rays as geometrical straight lines with zero diameter are an idealization, which is, however, very useful for the graphic construction of light propagation through optical systems.

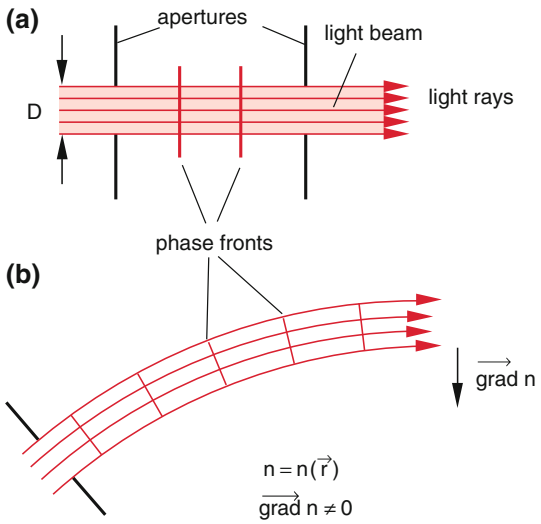


Fig. 9.1 Definition of light rays and light beams as laterally confined light waves. The normal to the phase front gives the propagation direction. **a**) In an optical homogeneous medium **b**) in optically inhomogeneous media with grad $n \neq 0$

The approximation of light rays has the following advantage: The investigation of the propagation of real waves through optical systems with many, often curved interfaces between media with different refractive indices (see Sect. 8.4) is very complicated. The approximation of geometrical optics allows a much simpler treatment which is for many applications sufficiently accurate.

In order to determine the propagation of light rays through optical instruments we have to introduce same basic facts of geometrical optics.

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_9

9.1 Basic Axioms of Geometrical Optics

The propagation of light waves follows some basic rules, which can be derived from theoretical principles as well as from experimental observations:

- In an optical homogeneous medium the light rays are straight lines.
- At the interface between two media light rays are reflected according to the reflection law (8.57) and they are refracted following Snelle's refraction law (8.58).
- Several light beams which intersect each other, do not influence each other if the intensities are not too high (region of linear optics). They do not deflect each other. In the superposition region interference effects can occur, but behind the intersection region the intensity distribution is not affected by the superposition.

Note, that this is no longer true for nonlinear optical phenomena.

The first two rules can be derived from ***Fermat's principle***, which was illustrated in Vol. 1. Sect. 11.11 for the refraction. It states, that light emitted from the point P_1 reaches the point P_2 always on such a path where the transit time is minimum. We will illustrate this by the example of reflection at a plane interface $y = 0$ (Fig. 9.2).

The path length from $P_1(x_1, y_1)$ via $R(x, 0)$ to $P_2(x_2, y_2)$ is

$$s = s_1 + s_2 = \sqrt{(x - x_1)^2 + y_1^2} + \sqrt{(x_2 - x)^2 + y_2^2}. \tag{9.1}$$

If the transit time $t = s/c$ should be minimum it follows

$$\begin{aligned} &\frac{dt}{dx} = 0 \\ &\Rightarrow \frac{x - x_1}{\sqrt{(x - x_1)^2 + y_1^2}} = \frac{x_2 - x}{\sqrt{(x_2 - x)^2 + y_2^2}} \\ &\Rightarrow \sin\alpha_1 = \sin\alpha_2. \end{aligned} \tag{9.2}$$

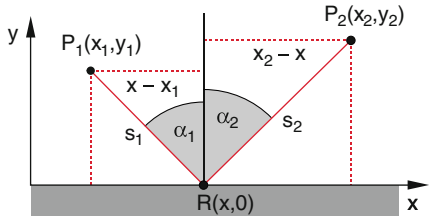


Fig. 9.2 Application of Fermat's principle to the reflection of a wave at a plane interface

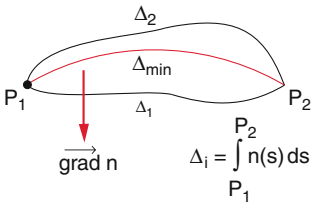


Fig. 9.3 Fermat's principle as variation principle for light beams in an optically inhomogeneous medium

The reflection law can be therefore deduced from Fermat's principle.

$$\sin\alpha_1 = \sin\alpha_2 \;\Rightarrow\; \alpha_1 = \alpha_2\,. \tag{9.3}$$

Fermat's principle is also valid in inhomogeneous media with locally changing refractive index. Here the light rays are curved (Fig. 9.1b). The principle of minimum transit time between the two points P_1 and P_2 is now (Fig. 9.3).

$$\delta \int_{P_1}^{P_2} n\,ds = 0\,, \tag{9.4}$$

where δ means an infinitesimal variation of the optical path length.

9.2 Optical Imaging

The goal of most optical arrangements is the generation of optical imaging, where the light emerging from a point P_1 is again concentrated in another point P_2. Such an imaging can be reached with a plane mirror, as can be seen in Fig. 9.4. Although all light rays emerging from P_1 are reflected divergently at the mirror plane, their extension into the lower half plane below the mirror plane intersect at the point P', the image point of P_1. An observer in the upper half plane sees the image P' behind the mirror. The image of an object

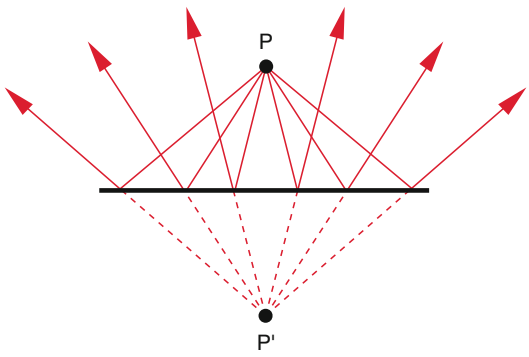


Fig. 9.4 Optical imaging by a plane mirror which produces from every arbitrary point above the mirror a virtual image below the mirror

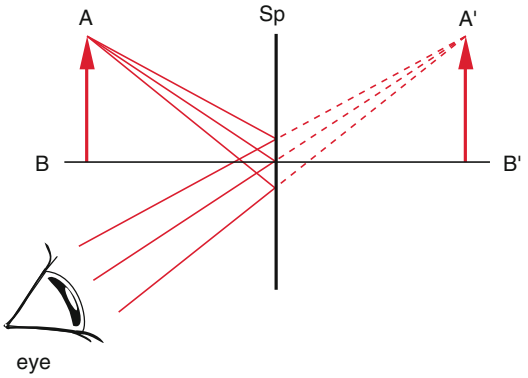


Fig. 9.5 A plane mirror images the object AB into the virtual image A' B' of the same size (Magnification $M = 1$)

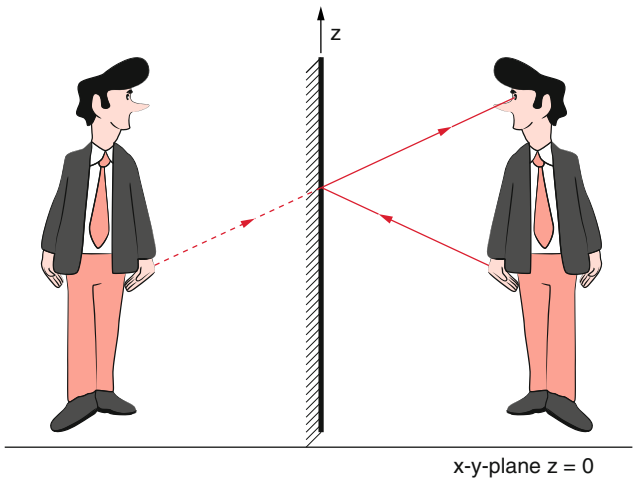


Fig. 9.6 Imaging by a plane mirror: person right, mirror image left. The image is laterally reversed. The right hand of the person becomes the left hand of the image

appears with the same size as the object itself (Fig. 9.5), it is, however, inversed left to right (Fig. 9.6) but not upside down!

The plane mirror is the only optical element that generates an ideal imaging. Each point P in space is imaged into a well defined other point P'.

There are other optical systems which image only selected points. One example is the elliptical mirror (Fig. 9.7) which images the two focal points into each other. A spherical mirror images only one point, the center M of the sphere into itself.

The approximate imaging of arbitrary points can be achieved with a very simple device, called the pinhole camera illustrated in Fig. 9.8. An illuminated or self-luminous object in the plane A is imaged through a small pinhole into the plane B. The diameter d of the pinhole can be varied. All light

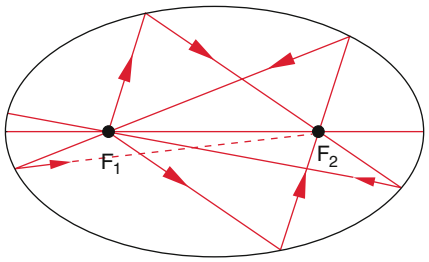


Fig. 9.7 An elliptical mirror images exactly two points (the focal points) into each other

rays starting from a point P are imaged into an elliptical area around the point P'. The larger diameter d' of this area is, according to theorem of intersecting lines

$$d' = \frac{a+b}{a} d \,. \tag{9.5}$$

This pinhole camera provides therefore no exact imaging of an arbitrary point P into the image point P' but into an area around P'. According to (9.5) the size of this area decreases with decreasing diameter d of the pinhole, i.e. the image becomes clearer but dimmer with decreasing pinhole diameter. This is illustrated in Fig. 9.9, which shows that there is an optimum diameter d, because for smaller values of d diffraction effects decrease the quality of the image (see Sect. 10.7.4). When the size $d_d = 2b \cdot \lambda/d$ of the central diffraction maximum exceeds the geometrical diameter $d' = d \cdot (a+b)/a$ in Fig. 9.8 the quality of the image becomes worse. The optimum diameter of the pinhole is therefore

$$d_{opt} = \sqrt{\frac{a \cdot b}{a+b} \cdot 2\lambda}. \tag{9.6}$$

Example

$\lambda = 500$ nm, $a = 20$ cm, $b = 5$ cm, => $d_{opt} = 0.2$ mm. Each point in the object plane A is the imaged into a circle with radius $r = 0.1$ mm around P'. This image sharpness is for many applications sufficient.

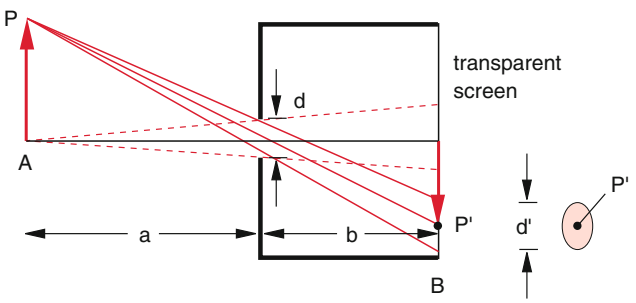


Fig. 9.8 Schematic representation of a pin hole camera

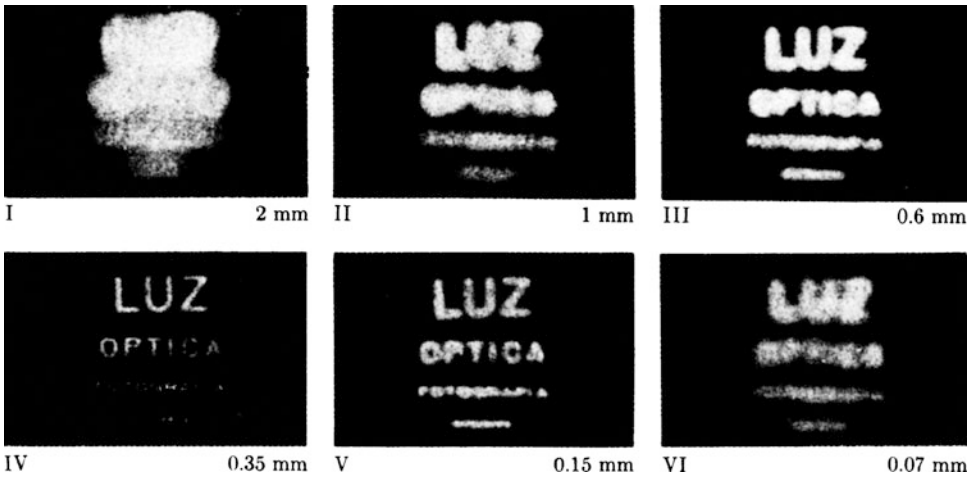


Fig. 9.9 Imaging of illuminated letters by a pin hole camera for different pin hole diameters. (Dr. N. Joel, Unesco, pilot project for the teaching of physics)

The main disadvantage of the pinhole camera is its small luminous intensity. The importance of imaging elements such as lenses or mirrors are the following:

- They allow much larger apertures and therefore transmit much higher light powers.
- They can produce the image of an object at every suitable distance.

Both points are very important for practical applications. However, all optical elements show imaging errors (see Sect. 9.5.6) which can be minimized by clever combination of different elements but which never can be completely eliminated. We will discuss this in the following by some examples.

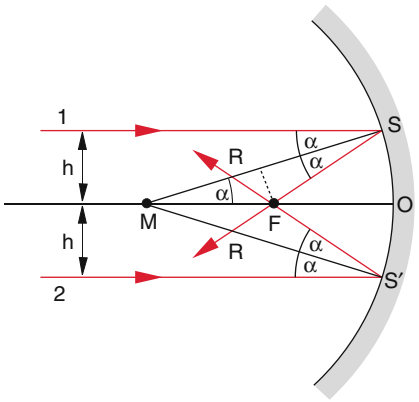


Fig. 9.10 Spherical mirror with radius R, focal point F, center M and focal length $f = OF \approx R/2$

9.3 Concave Mirrors

While plane mirrors generate distortion-free images of objects with a magnification 1:1, with curved mirrors enlarged or reduced images can be generated, which are, however, no longer distortion-free. We regard in Fig. 9.10 a spherical mirror with the center M. Two light rays 1 and 2 incident parallel to the mirror axis are reflected at the mirror surface according to the reflection law ($\alpha_i = \alpha_r = \alpha$). They intersect in the focal point F on the axis. The triangles MFS and MFS' in Fig. 9.10 are isosceles (since the two angles α are equal). Therefore the relation holds $FM = (R/2)/\cos\alpha$ and it follows

$$OF = R(1 - 1/(2\cos\alpha)). \tag{9.7a}$$

For sufficiently small distance h of the incident rays from the symmetry axis MO (paraxial rays) the angle α becomes small and we can approximate $\cos\alpha \approx 1$. In this case the focal length $f = OF$ becomes

$$f = R/2. \tag{9.7b}$$

For paraxial rays the focal length f of a spherical mirror equals half of its radius R.

Note The location of the focal point F depends on the distance h of the incident rays from the symmetry axis OM (Fig. 9.11)

With $\cos\alpha = \sqrt{(1 - \sin^2\alpha)}$ and $\sin\alpha = h/R$ we get for the focal length f

$$\begin{aligned} f &= R\left[1 - \frac{1}{2\cos\alpha}\right] \\ &= R\left[1 - \frac{R}{2\sqrt{R^2 - h^2}}\right]. \end{aligned} \tag{9.7c}$$

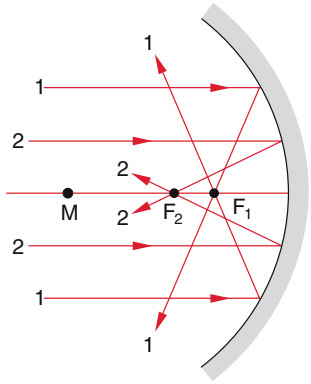


Fig. 9.11 The focal length of a spherical mirror is smaller for rays farther away from the axis than for rays closer to the axis

The focal length *f* of a spherical mirror decreases with increasing distance *h* of the incident rays from the symmetry axis.

Example

For $\alpha = 60°$ ($h = 0.87R$) the focal length $f = OF$ becomes $f = 0.3 \cdot R$

In Fig. 9.12 the image B of a point A at an arbitrary distance $g = OA > R$ is shown. We get for the angles shown in Fig. 9.12 the relations

$\delta = \alpha + \gamma$; (δ is exterior angle to the triangle BSM), and

$$\gamma + \beta = 2\delta. \tag{9.8}$$

For small angles γ (small distances h from the axis) we can use the approximations:

$$\gamma \approx \tan\gamma = \frac{h}{g},$$
$$\beta \approx \tan\beta = \frac{h}{b},$$
$$\delta \approx \sin\delta = \frac{h}{R},$$

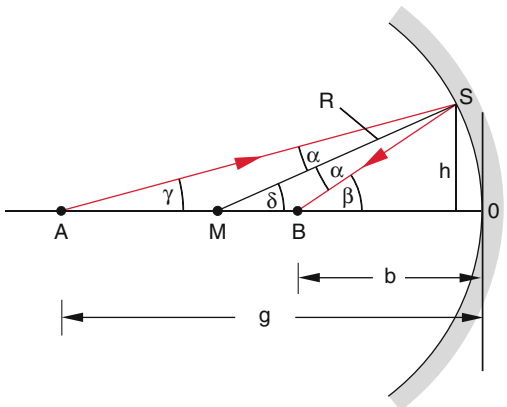


Fig. 9.12 Imaging of a point A on the symmetry axis into the image point B which lies also on the axis

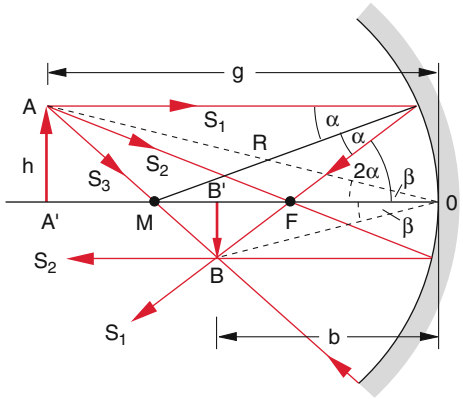


Fig. 9.13 Geometrical construction of the image B of an arbitrary point A close to the symmetry axis

With (9.8) and (9.7b) we then get the imaging equation

$$\frac{1}{g} + \frac{1}{b} \approx \frac{2}{R} \approx \frac{1}{f} \tag{9.9}$$

With the object distance g with the image distance b and the focal length f (9.9) is valid for incident rays with small values of h.

For the graphical construction of the image we regard in Fig. 9.13 the imaging of the arrow $A'A$ with the length h. We draw three rays starting rom A:

- The ray S_1 parallel to the symmetry axis MO which intersects after reflection the focal point F.
- The inclined ray S_2 which intersects F before the reflection and is therefore after reflection parallel to the axis MO
- The ray S_3 through the center M of the sphere which is reflected in itself.

All three rays intersect (in the approximation of paraxial rays with $h \ll f$) in point B, the image point of A. If the object distance $g = A'O$ is larger than the mirror radius $R = OM$, B is located between F and M but on the opposite side of the symmetry axis. The image B of A is reversed.

Remark For the graphical construction of the image B two rays would be sufficient. The third ray can be used to prove the consistency of the graphical construction.

The magnification factor BB'/AA' can be obtained from the relations

$$AA'/A'O = \tan\beta = BB'/B'O \Rightarrow BB'/AA' = B'O/A'O = b/g. \tag{9.10}$$

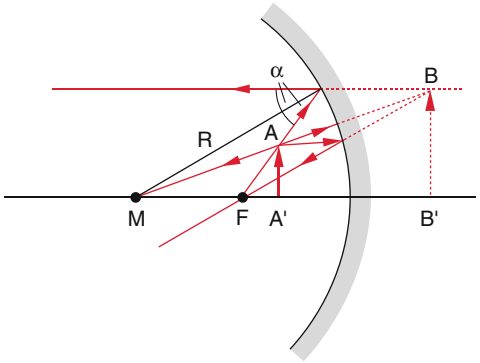


Fig. 9.14 Generation of a virtual image by a spherical mirror, if the object point A lies between mirror and focus F

The magnification factor is equal to the ratio of image distance b to object distance g.

When the object AA' is placed between mirror and focal point F the reflected rays are divergent (Fig. 9.14). Their opposite extensions intersect (in the paraxial approximation) in the point B behind the mirror. The image $\boldsymbol{BB'}$ is called a virtual image, because it is not a real image, and it cannot be seen on a screen placed at the position BB'. It just represents the mirror image of AA' seen by the eye.

When the center $\boldsymbol{M}$ of the curved mirror is on the same side as the object AA' the mirror is **concave** (Fig. 9.15a). If M and A lie on opposite sides of the mirror, the mirror is **convex** (Fig. 9.15b). A convex mirror can only produce virtual images produced on the other side of the mirror.

A special curved mirror, which is often used, in particular for headlights in cars and for astronomical telescopes, is the parabolic mirror (Fig. 9.16). A parabolic mirror focusses parallel light into the focal point F. It converts a plane wave into a nearly spherical wave. This can be seen from Fig. 9.16b and Fermat's principle.

The phase surfaces of the incident wave are the planes x = const. If all rays parallel to the symmetry axis (x-axis) intersect in the point F independent of their distance y from the x-axis, the optical path length from the plane

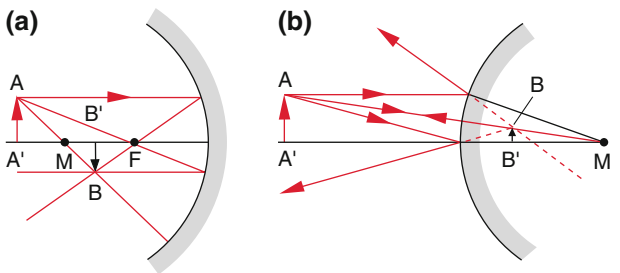


Fig. 9.15 **a**) Concave spherical mirror **b**) convex spherical mirror

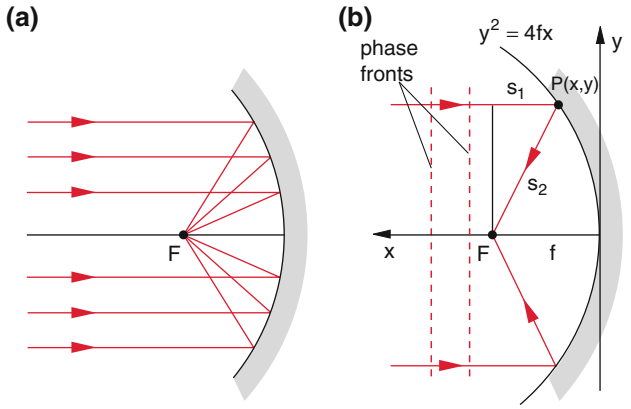


Fig. 9.16 **a**) Parabolic mirror **b**) application of Fermat's principle to the imaging of a plane parallel wave by a parabolic mirror

x = const. = f should be the same for all rays. The optical path length of a ray after reflection at the mirror point $P(x, y)$ is

$$s = s_1 + s_2 = f - x + \sqrt{(f-x)^2 + y^2}.$$

For $y^2 = 4fx$ the path length becomes $s = 2f$ independent of y. The equation of the mirror surface with the focal length f and the x-axis as symmetry axis is therefore

$$y^2 = 4fx \Rightarrow x = y^2/4f. \tag{9.11}$$

It is interesting to look for the difference between the parabolic and the spherical mirror. For the spherical surface in Fig. 9.17 we get instead of (9.11) the equation

$$\begin{aligned} y^2 + (R - x')^2 &= R^2 \\ \Rightarrow x' &= R - \sqrt{R^2 - y^2}. \end{aligned} \tag{9.12a}$$

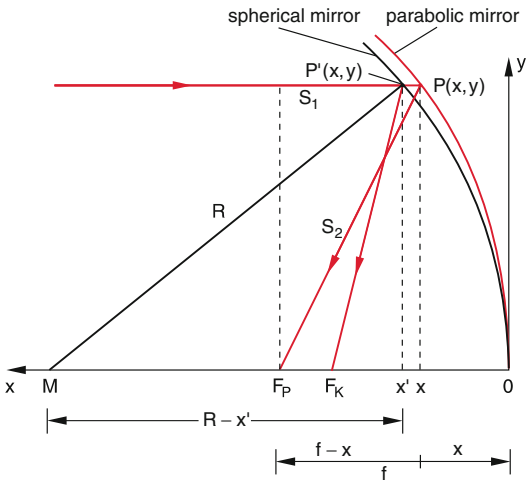


Fig. 9.17 Comparison of the ray paths for a spherical and a parabolic mirror with focal length $f = R/2$. For $y \ll R$ the focal point F_s of the spherical mirror moves towards that of the parabolic mirror F_P

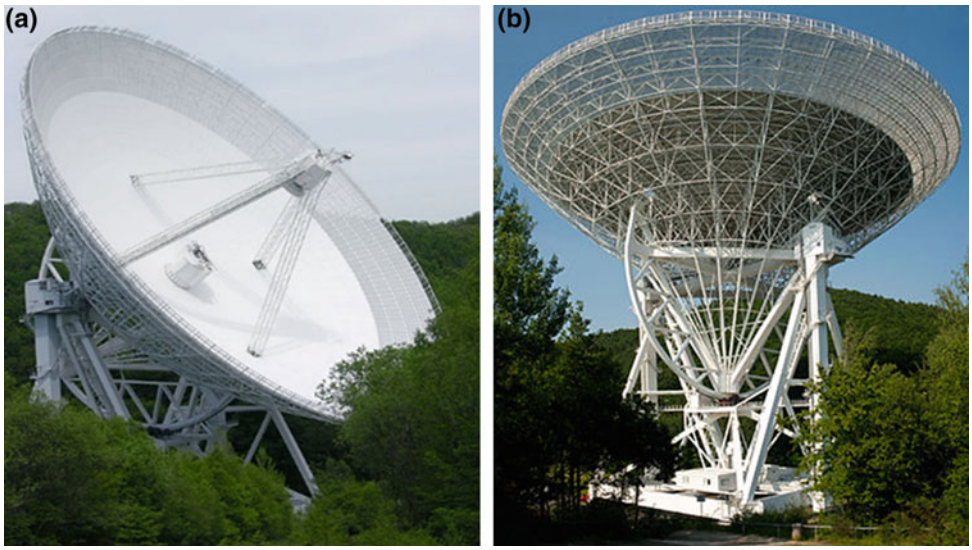


Fig. 9.18 Large radio-telescope in Effelsberg, Germany. The diameter of the parabolic mirror is 100 m its weight is 32,300 tons. It can be turned about a vertical and a horizontal axis, to cover a large angular section of the sky

For $y^2 < R^2$ the square root can be expanded into a Taylor series:

$$x = \frac{y^2}{2R} + \frac{y^4}{8R^3} + \frac{y^6}{16R^5} + \cdots . \tag{9.12b}$$

For paraxial rays the higher order terms can be neglected and we obtain with $f = R/2$ the Eq. (9.11) of a parabola. This shows:

> In the paraxial approximation the spherical mirror with radius R acts like a parabolic mirror with focal length $f = R/2$.

For rays farther away from the symmetry axis the focal length of the spherical mirror decreases while that of a parabolic mirror remains constant.

This means that the parabolic mirror can image incident beams with a larger diameter.

According to (9.12b) the distance $\Delta x = X(F_K) - X(F_P) \approx \frac{y^4}{8R^3}$ between the focal points of the spherical and the parabolic mirror increases with the distance y of the incident rays proportional to y^4.

Note that the parabolic mirror has the same focal point for all rays incident parallel to the symmetry axis, while for the spherical mirror this is only true for paraxial rays, i.e. rays with small distances y from the symmetry axis.

Parabolic mirrors are used in astronomy. One example is the large mirror in Effelsberg, Germany, with a diameter of 100 m (Fig. 9.18). It is used for receiving radio signals from the universe and it can be rotated and tilted in order to reach a large angular range in the sky. The radio radiation received by the parabolic mirror is focused onto a detector cooled to a low temperature of about 10 K. Radiation with a wavelength of $\lambda = 21$ cm is emitted by hydrogen atoms in our galaxy on a hyperfine transition. Also rotational transitions in molecules can be detected by the telescope and many molecules have been found in space up to large biological molecules.

9.4 Prisms

A light ray passing through an isosceles prism is two times refracted and its total deflection δ against the incident beam is, according to Fig. 9.19.

$$\delta = \alpha_1 - \beta_1 + \alpha_2 - \beta_2 .$$

We can express the deflection angle δ by the incident angle α_1 and the prism angle γ. From Fig. 9.19 we can derive the relations $\gamma = \beta_1 + \beta_2$ (because the sum off the three angles in the triangle ABC is

$$\begin{aligned} \gamma + (90^\circ - \beta_1) + (90^\circ - \beta_2) = 180^\circ \Rightarrow \\ \delta = \alpha_1 + \alpha_2 - \gamma . \end{aligned} \tag{9.13}$$

Minimum deflection at a fixed angle γ occurs, if $d\delta/d\alpha_1 = 0$

This gives

$$\frac{d\delta}{d\alpha_1} = 1 + \frac{d\alpha_2}{d\alpha_1} = 0 \Rightarrow d\alpha_2 = -d\alpha_1 . \tag{9.14}$$

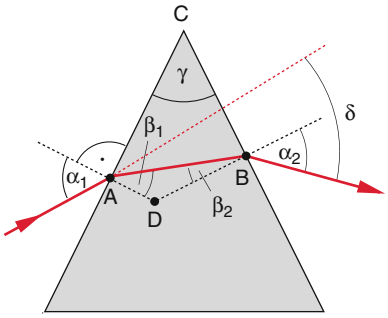


Fig. 9.19 Deflection of a light ray in a prism

From the derivatives of Snell's refraction law $\sin\alpha = n \cdot \sin\beta$ we get

$$\cos\alpha_1 \mathrm{d}\alpha_1 = n \cdot \cos\beta_1 \cdot \mathrm{d}\beta_1, \tag{9.15a}$$

$$\cos\alpha_1 \mathrm{d}\alpha_1 = n \cdot \cos\beta_2 \cdot \mathrm{d}\beta_2. \tag{9.15b}$$

From $\beta_1 + \beta_2 = \gamma$ and $d\gamma/d\alpha\ 1 = 0$ (because γ = const.) we can deduce $\mathrm{d}\beta_1 = -\mathrm{d}\beta_2$. Dividing (9.15a) by (9.15b) gives

$$\frac{\cos\alpha_1}{\cos\alpha_2}\frac{\mathrm{d}\alpha_1}{\mathrm{d}\alpha_2} = \frac{\cos\beta_1}{\cos\beta_2}.$$

For the ray passing with minimum deflection δ ($\mathrm{d}\alpha_1 = -\mathrm{d}\alpha_2$) this can be reduced to

$$\frac{\cos\alpha_1}{\cos\alpha_2} = \frac{\cos\beta_1}{\cos\beta_2},$$

This can be transformed, using the refraction law, into

$$\frac{1-\sin^2\alpha_1}{1-\sin^2\alpha_2} = \frac{n^2-\sin^2\alpha_1}{n^2-\sin^2\alpha_2}. \tag{9.16}$$

Since $n \neq 1$ this can be only fulfilled if $\alpha_1 = \alpha_2 = \alpha$.

For the symmetric ray path with $AC = BC$ and $\alpha_1 = \alpha_2$ the deflection δ is minimum. For the incident angle α the total deflection δ of rays passing through an isosceles prism with prism angle γ is

$$\delta_{\min} = 2\alpha - \gamma. \tag{9.17}$$

With the refraction law $\sin\alpha = n \cdot \sin\beta$ one obtains the relation

$$\sin\frac{\delta_{\min}+\gamma}{2} = \sin\alpha = n \cdot \sin\beta = n \cdot \sin(\gamma/2). \tag{9.18}$$

The dependence of δ on the refractive index n can be derived from (9.18) with $\mathrm{d}\delta/\mathrm{d}n = (\mathrm{d}n/\mathrm{d}\delta)^{-1}$. The result is

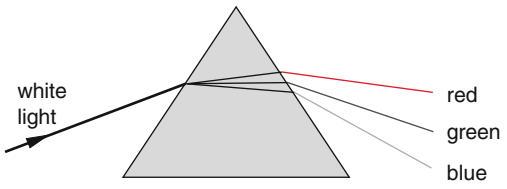


Fig. 9.20 Within the spectral range of normal dispersion ($\mathrm{d}n/\mathrm{d}\lambda < 0$) blue light is deflected more than red light

$$\frac{\mathrm{d}\delta}{\mathrm{d}n} = \frac{2\sin(\gamma/2)}{\cos[(\delta+\gamma)/2]} = \frac{2\sin(\gamma/2)}{\sqrt{1-n^2\sin^2(\gamma/2)}}. \tag{9.19}$$

Since the refractive index $n(\lambda)$ depends on the wavelength λ (dispersion, Sect. 8.2) we finally arrive at the relation between δ and λ using $\mathrm{d}\delta/\mathrm{d}\lambda = (\mathrm{d}\delta/\mathrm{d}n) \cdot (\mathrm{d}n/\mathrm{d}\lambda)$.

$$\frac{\mathrm{d}\delta}{\mathrm{d}\lambda} = \frac{2\sin(\gamma/2)}{\sqrt{1-n^2\sin^2(\gamma/2)}} \cdot \frac{\mathrm{d}n}{\mathrm{d}\lambda}. \tag{9.20}$$

Figure 9.20 illustrates the deflection of a parallel white light beam which is separated into the different colors because of the wavelength-dependent refractive index $n(\lambda)$.

For most transparent media is in the visible range $\mathrm{d}n/\mathrm{d}\lambda < 0$ (normal dispersion). This implies that blue light experience a larger deflection than red light.

Example

For an isosceles prism with ($\gamma = 60°$) is

$$\frac{\mathrm{d}\delta}{\mathrm{d}\lambda} = \frac{\mathrm{d}n/\mathrm{d}\lambda}{\sqrt{1-n^2/4}}.$$

With $\mathrm{d}n/\mathrm{d}\lambda = 4 \times 10^5\ \mathrm{m}^{-1}$ at the wavelength $\lambda = 400$ nm and $n = 1.8$ (for Flint glass) we obtain $\mathrm{d}\delta/\mathrm{d}\lambda = 1 \times 10^3$ rad/nm. Two wavelengths λ_1 and λ_2 which differ by $\Delta\lambda = 10$ nm experience deflection angles that differ by 10^{-2} rad $\approx 0.6°$.

9.5 Lenses

Optical lenses had an enormous influence onto the development of optics over the last centuries. The lens maker *Hans Lipershey* (1570–1619) constructed in Holland the first telescope with lenses that he had grinded himself. Copying and essentially improving this first telescope Galilei could observe 1610 for the first time the four largest moons of Jupiter (Io, Europe, Ganymede and Calisto, called the Galilean moons, see Vol. 1, Fig. 1.1).

Besides the telescope many other optical instruments (e.g. spectacles, magnifying glass, microscope, projectors cameras) are based on optical lenses (see Chap. 11). It is therefore worthwhile to study the optical characteristics of lenses in more detail.

9.5.1 Refraction at a Curved Surface

We regard in Fig. 9.21 an optical ray parallel to the symmetry axis with a distance h, which impinge onto a spherical interface between two media with refractive indices n_1 and n_2. The ray is refracted at the point A on the surface, propagates on a straight line in the homogeneous medium and intersects the symmetry axis in the focal point F. From Fig. 9.21 we can derive

$$h = R \cdot \sin\alpha = f \cdot \sin\gamma.$$

With $\gamma = \alpha - \beta$ we get the focal length f

$$f = \frac{\sin\alpha}{\sin(\alpha - \beta)} \cdot R.$$

With the refraction law

$$n_1 \cdot \sin\alpha = n_2 \cdot \sin\beta$$

we obtain with $\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\sin\beta$ for small angles ($\cos\alpha \approx \cos\beta \approx 1$) the focal length

$$f = \frac{n_2}{n_2 - n_1} \cdot R. \qquad (9.21a)$$

Example

For the interface between air ($n_1 = 1$) and glass ($n_2 = 1.5$) Eq. (9.21a) gives $f = 3R$. For $n_1 = 1$ and $n_2 = 3 \Rightarrow f = 1.5R$

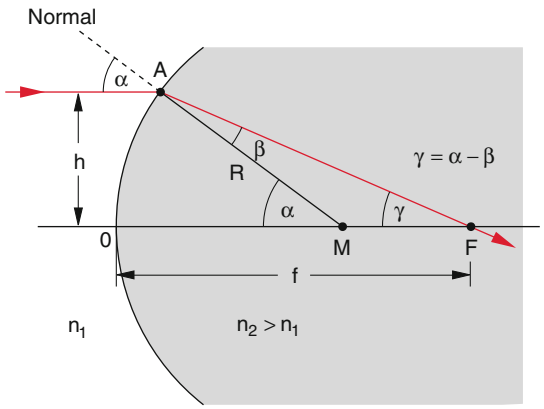


Fig. 9.21 Definition of the focal length of a spherical curved surface

Note Equation (9.21a) is only valid for paraxial rays ($h \ll R$).

Analogue to the construction of the image by a curved mirror the image B of an object A can be graphically constructed, by drawing at least two rays (Fig. 9.22).

The ray parallel to the symmetry axis, which passes through the focal point F_2 and the ray through the center point M of the curved surface, which passes without refraction through the interface. The two beams intersect at the image point B.

A third beam can be used for checking the accuracy of the first two rays, which passes through the left focal point F_1 and propagates in the second medium parallel to the symmetry axis.

Of course one can also construct the reverse propagation, regarding the point B as the object and A as the image point. The ray from B parallel to the axis in medium 2 intersects the symmetry axis in medium 1 at the focal point F_1 on the object side. One obtains for the focal length on the object side

$$f_1 = \left(\frac{n_1}{n_1 - n_2}\right) R. \qquad (9.21b)$$

If the object A has the distance a from the point O (Fig. 9.23) we can deduce from the approximate refraction law $n_1 \cdot \alpha \approx n_2 \cdot \beta$ for paraxial rays with $\alpha = \delta + \varepsilon$ and $\beta = \delta - \gamma$ (δ and α are exterior angles to the triangle APM, resp. PMB) the relations

$$n_1(\delta + \varepsilon) \approx n_2(\delta - \gamma). \qquad (9.22a)$$

The distance PX in Fig. 9.23 can be expressed for paraxial rays as

$$\begin{aligned} PX &= (a + x)\tan\varepsilon \approx a \cdot \varepsilon, \text{weil } x \ll a, \tan\varepsilon \\ &= (b + x)\tan\gamma \approx b \cdot \gamma \\ &= R \cdot \sin\delta \approx R \cdot \delta. \end{aligned}$$

Inserting this into (9.22a) yields, after rearrangement and division by PX, the relation

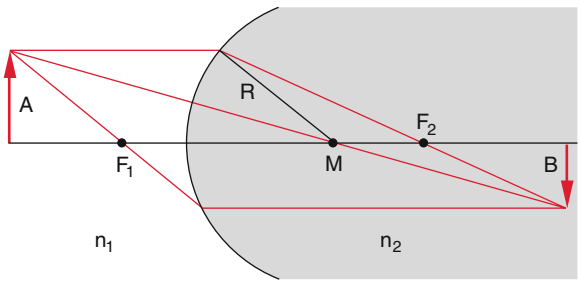


Fig. 9.22 Geometrical construction of light rays for the imaging of an object A by a spherical surface into the image B

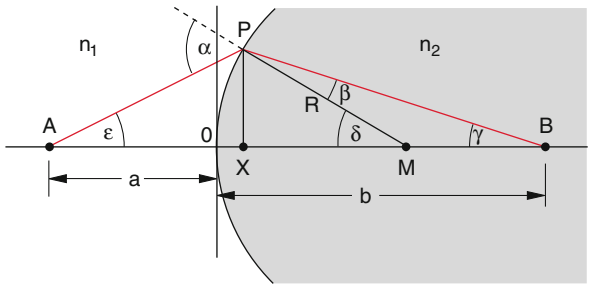


Fig. 9.23 Illustration of Eq. (9.22a–9.22c)

$$\frac{n_1}{a} + \frac{n_2}{b} = \frac{n_1 - n_1}{R} \tag{9.22b}$$

between the object distance a, the image distance b and the radius of curvature R. Using (9.21a and 9.21b) we can express this by the focal length f and obtain

$$\frac{n_1}{a} + \frac{n_2}{b} = \frac{n_2}{f_2} = -\frac{n_1}{n_2}. \tag{9.22c}$$

9.5.2 Thin Lenses

A lens consists of a transparent medium with refractive index n_2 which is separated on both sides by polished surfaces from a medium with refractive index n_1 (generally air with $n_1 = 1$) (Fig. 9.24).

We will here restrict the discussion on lenses with spherical surfaces in air. We can then set $n_1 = 1$ and $n_2 = n$. The different types of lenses are classified according to their radii of curvature R_1 and R_2 which are defined as the orientated distance from the curved surface to its center of curvature M (Fig. 9.25). The radius of curvature is positive ($R > 0$) if it points into the positive direction to the right of the curved surface, it is negative ($R < 0$) if it points into the negative direction.

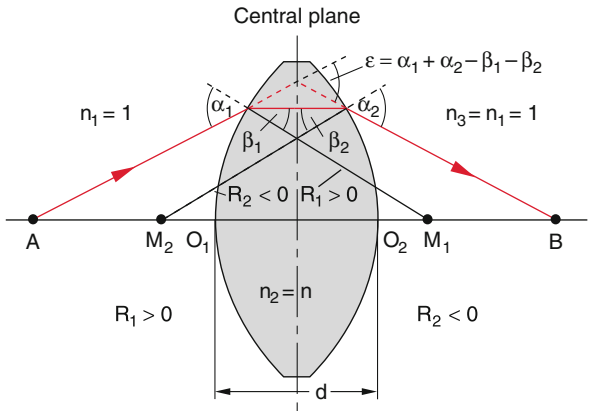


Fig. 9.24 Imaging of the object point A on the symmetry axis into the image B by a lens with radii R_1 and R_2

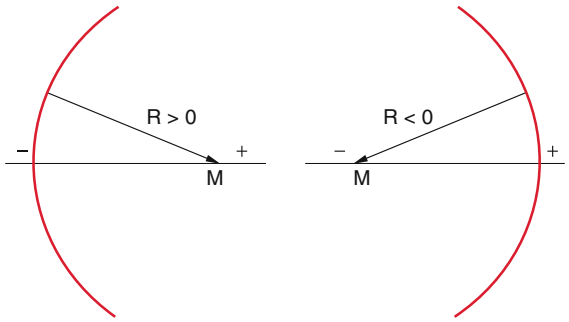


Fig. 9.25 Definition of the signs of radii of curvature

Note We will always use the convention that the incident light propagates from left to right, i.e. from the negative to the positive half space (Fig. 9.24). We can therefore also use the equivalent definition that R is positive if the center of curvature M lies on that side of the interface which is opposite to the light source.

The interface in Fig. 9.23 has, for instance a positive radius of curvature. In Fig. 9.24 is $R_1 > 0$ and $R_2 < 0$.

In Fig. 9.26 some types of lenses are illustrated. A curved lens surface is **convex,** if the lens lies between surface and the center M of curvature, otherwise it is **concave**. The types (a), (b) and (f) in Fig. 9.26 are convergent lenses, (d) and (e) are diverging lenses. The form (c) is a convergent lens for $|R_1| < |R_2|$ but a diverging lens otherwise.

A **thin lens** is the idealization of real lenses where the maximum distance between the two surfaces is small compared to the focal length.

The optical imaging by a lens can be described by successive imaging by the two surfaces of the lens (Fig. 9.27). For the first surface we obtain from (9.22a) to (9.22c) with $n_1 = 1$ and $n_2 = n$

$$\frac{1}{a_1} + \frac{n}{b_1} = \frac{n-1}{R_1}. \tag{9.23a}$$

If only the interface 1 with radius of curvature R_1 would exist (i.e. to the right of the surface 1 extends only the homogeneous medium with refractive index n_2) the point A would be imaged into the point B_1 in Fig. 9.27a.

By the second refraction at the second interface the rays are again bent and intersect in the image point B which is closer to the lens. Equation 9.22b for the second imaging can be obtained by the following consideration: When we reverse the role of object and image we can regard B as the object which is imaged into the point A. Since we have now reversed the direction of imaging we also have to reverse the succession of

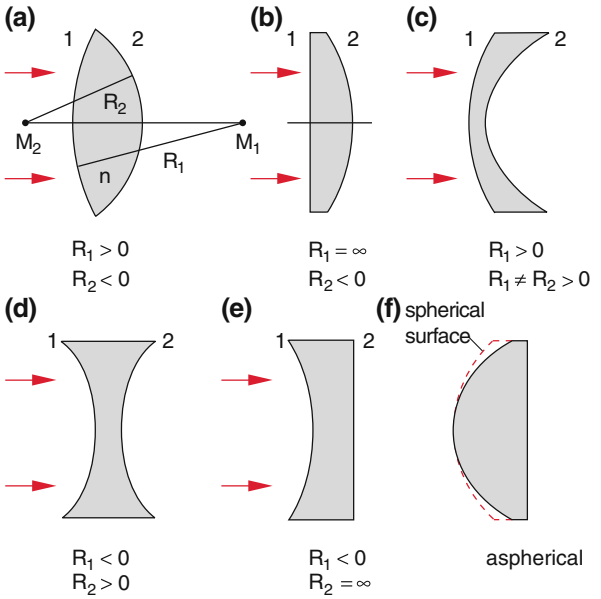


Fig. 9.26 Examples of different forms of lenses: **a**) convex-convex = biconvex **b**) plane-convex **c**) convex-concave **d**) biconcave **e**) concave-plane **f**) aspherical lens

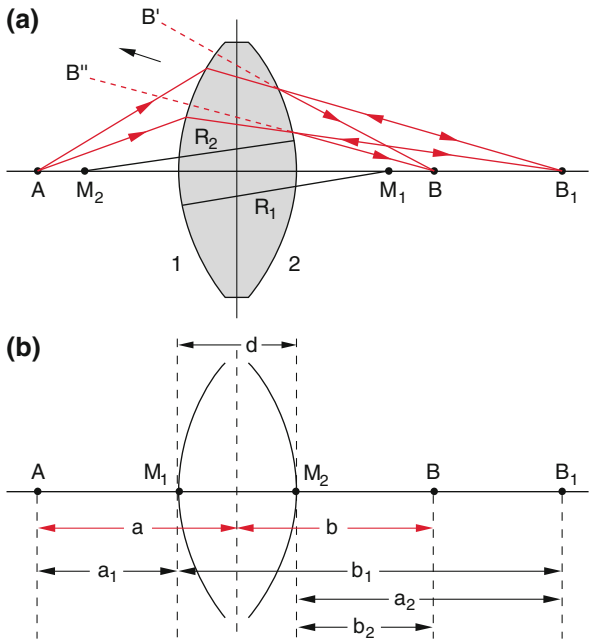


Fig. 9.27 Illustration of the derivation of the lens equation

the refractive indices. The rays starting from B in the reverse direction would proceed, without refraction at the surface 2, along the dashed lines. An observer in B would assume that the light rays come from the points B' or B''. For considering the imaging by the curved surface 2 we have to set (Fig. 9.27b) $n_1 \rightarrow n$, $n_2 \rightarrow 1$, $a \rightarrow -(b_1 - d)$, and $R \rightarrow -R_2$. The minus sign appears because the imaging now proceeds from right to left. We then obtain

$$\frac{-n}{b_1 - d} + \frac{1}{b_2} = \frac{1-n}{R_2}. \tag{9.23b}$$

Adding (9.23a) and (9.23b) gives the equation

$$\frac{1}{a_1} + \frac{1}{b_2} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right) + \frac{n \cdot d}{b_1(b_1 - d)}. \tag{9.24a}$$

Introducing the distances $a = a_1 + d/2$ and $b = b_2 + d/2$ from A until the mid of the lens we obtain for thin lenses the **lens equation**

$$\frac{1}{a} + \frac{1}{b} = (n-1)\left(\frac{1}{R_1} - \frac{1}{R_2}\right). \tag{9.24b}$$

This is the general equation for the imaging by thin lenses, where the distance O_1O_2 is small compared to the focal lengths f_1 and f_2. For the graphic construction one can replace the refraction at the two lens surfaces by a single refraction at the center plane of the lens with the refraction angle $(\alpha_1 - ß_1) + (\alpha_2 - ß_2)$ (Figs. 9.27 and 9.28).

For an incident beam parallel to the axis is in (9.24b) $a = \infty$. Since this ray has to pass through the focal point F is $b = f$ and we get for the focal length of a thin lens

$$f = \frac{1}{n-1}\left(\frac{R_1 \cdot R_2}{R_2 - R_1}\right). \tag{9.25a}$$

For a biconvex lens with equal radii of curvature ($R_1 = -R_2 = R$) the focal length becomes

$$f = \frac{R/2}{n-1}. \tag{9.25b}$$

Compare this result with the focal length $f = R/2$ of a spherical mirror.

Inserting the focal length (9.25a) into (9.24a and 9.24b one obtains the imaging equation of thin lenses

$$\frac{1}{a} + \frac{1}{b} = \frac{1}{f}. \tag{9.26}$$

For the graphic construction of the imaging by thin lenses one uses the ray 1 parallel to the axis (Fig. 9.28) which passes through the focal point F_2 on the image side and the ray 2 through the center point O of the lens, which is not deflected. The displacement Δ of this ray

$$\Delta = d \cdot \sin\alpha\left(1 - \frac{\cos\alpha}{\sqrt{n^2 - \sin^2\alpha}}\right)$$

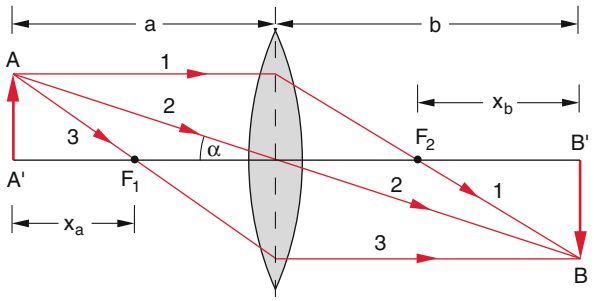


Fig. 9.28 Geometrical construction of the imaging by a thin lens

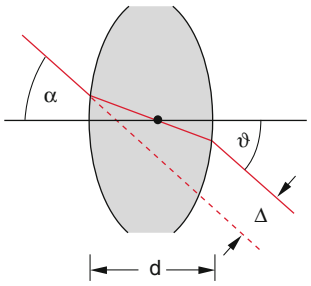


Fig. 9.29 The ray passing through the center of a lens is not deflected but displaced by the distance Δ

can be neglected for thin lenses with $d \to 0$. Assuming B as object point and A as image point the ray 3 parallel to the axis on the image side has to pass through F_1 on the object side. Inserting in (9.26) $a = f + x_a$ and $b = f + x_b$ we obtain for the distances x_a between object point A and focal point F_1 and x_b between image point B and focal point F_2 Newton's imaging equation

$$x_a \cdot x_b = f^2. \tag{9.26a}$$

With the lateral magnification (imaging scale….) M this yields

$$M = \frac{\overline{BB'}}{\overline{AA'}}$$

The quantity M can be immediately obtained from Fig. 9.28 using the relation

$$M = -\frac{b}{a} = \frac{f}{f-a}. \tag{9.27}$$

For $M < 0$ the image of the object is reversed, for $M > 0$ the image arrow in Fig. 9.28 has the same direction as the object arrow.

One can see from (9.27) that $M < 0$ for $a > f$. This means: The imaging is always reversed (the image arrow has the opposite direction as the object arrow) if the object distance a is larger than the focal length f. For a = $2f \to M = -1$, i.e. object and image have the same size but opposite directions. For $a = f \to b = \infty$, i.e. the image is shifted towards infinity and $M = \infty$.

9.5.3 Thick Lenses

For thin lenses we could replace the twofold refraction at the two lens surfaces by a single refraction at the center plane of the lens. For thick lenses where the distance S_1S_2 of the vertices of the two surfaces is no longer negligible; this simplification would lead to larger deviations from the real situation. Looking at the ray passing through the center O of the thick lens (Fig. 9.29) one can see that the output ray is not deflected but only displaced by the displacement Δ. This offers the following substitute of the ray path through thick lenses (Fig. 9.30): The incident beam is prolonged up to its intersection P_1 with the axis. Then it proceeds along the axis until the point P_2 where it follows the straight extension of the exit ray. The vertical planes through the points P_1 and P_2 are called the **principal planes**. The refraction of rays at the surfaces of thick lenses can be replaced by two refractions at the principles planes (instead at one plane for thin lenses).

This construction replaces the thick lens by two thin lenses located at the principle planes through P_1 and P_2. with their distance h. It is possible to prove with some mathematical efforts that one obtains also for thick lenses the imaging Eq. (9.26) of thin lenses, if the objet distance is measured from the object A to the point P_1 of the first principal plane, and the image distance from P_2 to the image point B. For the focal length of a thick lens with the thickness $d = h_1 + h + h_2$ in air one obtains instead of (9.25a) the expression

$$\frac{1}{f} = (n-1)\left[\frac{1}{R_1} - \frac{1}{R_2} + \frac{(n-1)d}{nR_1R_2}\right]. \tag{9.28}$$

For the distances between the vertices S_i and the points P_i one gets

$$h_1 = -\frac{(n-1)f \cdot d}{n \cdot R_2}, \qquad h_2 = -\frac{(n-1)f \cdot d}{n \cdot R_1}, \tag{9.29}$$

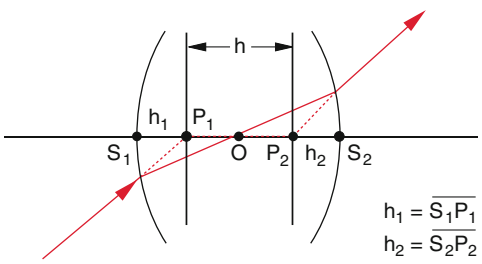


Fig. 9.30 Definition of the principle planes P_1 and P_2 of a thick lens

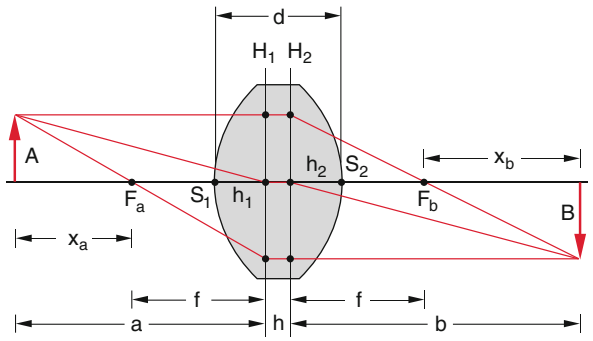


Fig. 9.31 Geometrical construction for the imaging by a thick lens

where $h_i > 0$ if P_i is located on the right side of $S_i(x(P_i) > x(S_i))$ and $h_i < 0$ if P_i is left of S_i. Note that the signs of f and R_i have to be considered according to the conventions in Fig. 9.25.

The intersections P_i of the principle planes with the symmetry axis are called **principal points** of the lens. For $d \to 0$ the principle planes converge against the center plane of the thin lens and (9.28) transfers into (9.25a).

In Fig. 9.31 some examples of the principle planes of different forms of thick lenses are shown. This illustrates that the principle planes can indeed lie outside the lens. The principle planes are often designated by the letter H (from the German Hauptebene).

Example

For a biconvex lens with $N = 1.5$, $R_1 = 20$ cm, $R_2 = -30$ cm and $d = 1$ cm the focal length becomes according to (9.28) $f = 24$ cm. The principle planes have the distances $h_1 = +2.6$ cm rom S_1 and $h_2 = -4.0$ from S_2 as can be proved by inserting the numbers into (9.29) (Fig. 9.32).

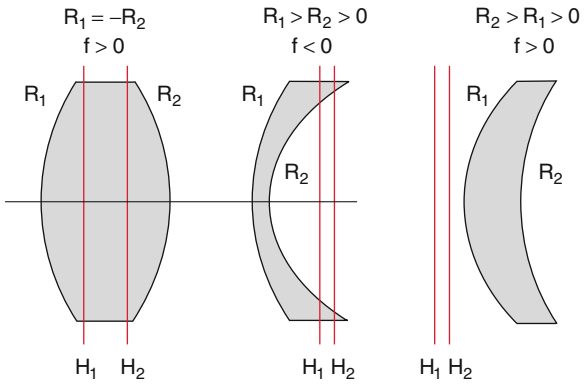


Fig. 9.32 Examples of the position of the principle planes for different forms of lenses

The geometrical construction of the imaging by a thick lens is analogue to that for a thin lens (Fig. 9.28), if the principle planes are regarded as the refracting planes (Fig. 9.31). The object distance a is measured from A to P_1, the image distance b from P_2 to B (Fig. 9.30).

The distances x_a between object A and focal point F_a and x_b between focal point F_b and image B can be determined from Fig. 9.31 using the theorem of intersecting lines

$$\frac{x_a}{f} = \frac{A}{B} \text{ and } \frac{x_b}{f} = \frac{B}{A} \Rightarrow x_a \cdot x_b = f^2. \tag{9.30}$$

With $x_a = a - f$ and $x_b = b - f$ this gives the lens Eq. (9.26)

$$f = \frac{a \cdot b}{a + b} \Rightarrow \frac{1}{f} = \frac{1}{a} + \frac{1}{b} \tag{9.31}$$

which is valid also for thick lenses. The only difference is that a and b are measured from the principal planes P_1 and P_2 and not, as for thin lenses, from the center plane of the lens.

9.5.4 System of Lenses

Often more than one lens has to be used for special imaging problems (see Sects. 9.5.5 and Chap. 11). The optimum choice for the combination of several lenses can considerably improve the quality of the image. We will illustrate for the example of two lenses the method for determining the relevant parameters of a lens system.

We regard in Fig. 9.33 a system of two thick lenses L_1 and L_2 with focal lengths f_1 and f_2 and the distance $D = P_{12}P_{21}$ between the inner principal planes P_{12} and P_{21}. A ray from the object A parallel to the axis passes through the focal point F_{b1} on the image side of L_1 and propagates further on until it reaches the focal point F_b of the lens system. An object A at a far distance ($a \to \infty$) is imaged by L_1 into its focal plane $b = f_1$.

The intermediate image of L_1 in F_{b1} has for L_2 the object distance $a_2 = D - f_1$ is further imaged by L_2 according to (9.31) into the image distance

$$b_2 = \frac{a_2 f_2}{a_2 - f_2} = \frac{(D - f_1) f_2}{(D - f_1 - f_2)}. \tag{9.32a}$$

For arbitrary distances a_1 it follows from (9.31) $b_1 = \dfrac{a_1 f_1}{(a_1 - f_1)}$.

The focal length f of the total system can be defined as (see problem 9.15)

$$f = \frac{f_1 \cdot f_2}{f_1 + f_2 - D} \tag{9.32b}$$

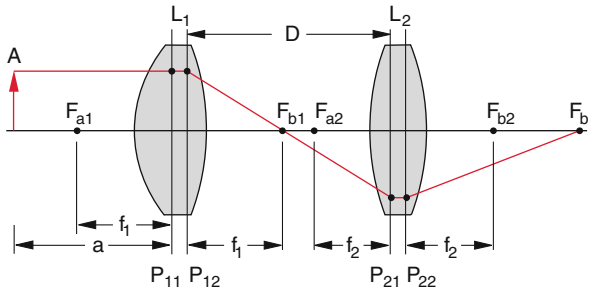


Fig. 9.33 Example of an optical system of two thick lenses

This gives for the lens system the imaging equation analogue to the Eq. (9.31) for a singel lens

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{D}{f_1 f_2}, \tag{9.32c}$$

where the focal lengths f_i are defined as shown in Fig. 9.33. For $D << f_1$ and $D << f_2$ we can neglect the last term and obtain the result:

The reciprocal focal length of two close lenses add up to the reciprocal focal length of the lens system.

The reciprocal focal length $D^* = 1/f$ of a lens is called its **refractive power**. It is measured in units of **diopter** D^*, where 1dpt = 1 m^{-1}.

Equation (9.32c) can be formulated as

The refractive powers of two close lenses centered around the same symmetry axis add up to the total refractive power of the lens system.

Example

A lens with $f = 50$ cm has a refractive power $D_1^* = 1/(0{,}5\ \text{m}) = 2$ dpt. A lens L_2 with $f_2 = 30$ cm has a refractive power $D_2^* = 3.33$ dpt. The total system has the refractive power $D^* = D_1^* + D_2^* = 5.33$ dpt and therefore a focal length $f = 18.8$ cm if the distance D between the two lenses can be neglected.

Choosing the right combination of f_1, f_2 and D in (9.32c) any wanted focal length of the system can be realized.

Example

Two lenses with $f_1 = 20$ cm and $f_2 = 30$ cm and the distance D give the focal length of the system

$$f = \frac{20 \cdot 30}{20 + 30 - D}\ \text{cm}.$$

For $D < 50$ cm is $f > 0$, the system acts as collecting lens. For $D > 50$ cm the total focal length becomes negative The system acts as diverging lens. For $D = 6$ cm the focal length is $f = 13.6$ cm, for $D = 60$ cm $\rightarrow f = -60$ cm.

Figures 9.33 and 9.34 show the geometric construction of two different lens systems where in Fig. 9.34 the distance $D > f_1 + f_2$ is larger than the sum of the two focal lengths, whereas in Fig. 9.35 $D < f_1$ and f_2 is smaller than each of the two focal lengths. In Fig. 9.35 the image B is inverted. Without

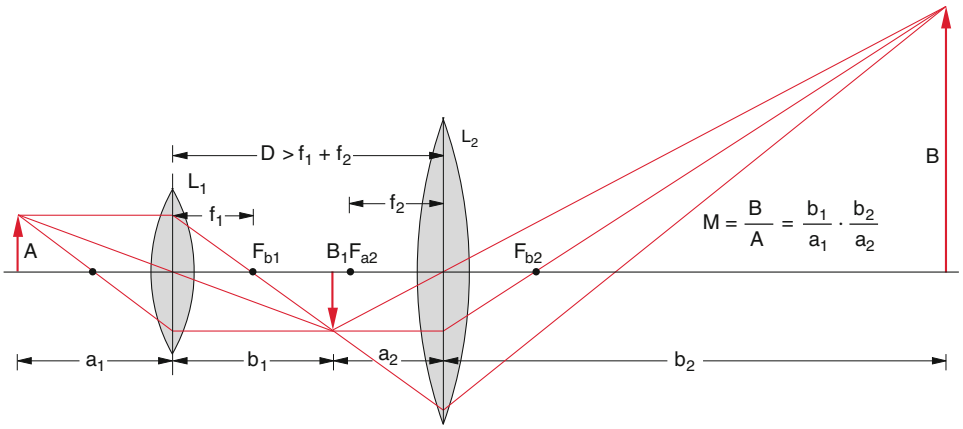


Fig. 9.34 Imaging by a system of two lenses with a distance $D > f_1 + f_2$. The intermediate image B_1 produced by L_1 is further imaged by L_2 into the final image B

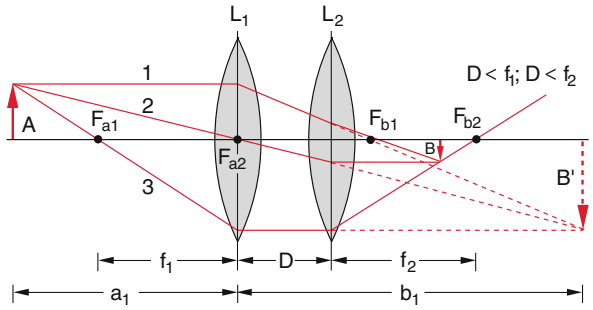


Fig. 9.35 Imaging by a system of two lenses with a distance $D < f_1$ and $D < f_2$

the lens L_2 the larger image B' would be generated (by the dashed straight lines). Due to the refraction by the second lens L_2 the three ray intersect at the smaller image B.

The magnification factor M of the lens system is for $D > f_1 + f_2$ equal to the product $M_1 \cdot M_2$ of the two lenses. From Fig. 9.34 we obtain

$$M = M_1 \cdot M_2 = \frac{b_1}{a_1} \cdot \frac{b_2}{a_2} = \frac{b_1 b_2}{a_1 (D - b_1)}, \tag{9.33a}$$

because $a_2 = D - b_1$. Using the expression (9.27) for the magnification factors M_1 and M_2 one obtains

$$\begin{aligned} M &= \frac{f_1 \cdot f_2}{(f_1 - a_1)(f_2 + b_1 - D)} \\ &= \frac{1}{(1 - a_1/f_1)(1 + (b_1 - D)/f_2)}. \end{aligned} \tag{9.33b}$$

Replacing b_1 with the expression in the imaging Eq. (9.26) the result of the magnification factor becomes

$$M = \frac{1}{1 - \frac{a_1}{f_1} - \frac{a_1 + D}{f_2} + \frac{a_1 D}{f_1 f_2}}. \tag{9.33c}$$

More information about systems of several lenses can be found, for instance, in [1].

9.5.5 Zoom-Lens Systems

For many applications, in particular for photo-shooting or video cameras a variable magnification without changing the lenses is very useful. This can be achieved by system of lenses, where the distance D between the lenses can be varied, which changes the magnification M without changing the object plane or the image plane (see Eq. (9.33c)). Such lens systems are called **Zoom lenses**. They consist of at least 3 lenses. In Fig. 9.36 a system of 4 lenses is shown consisting of a pair of divergent lenses with a firmly fixed distance between two fixed converging lenses. The pair can be shifted by the distance d^*. The magnification M of the

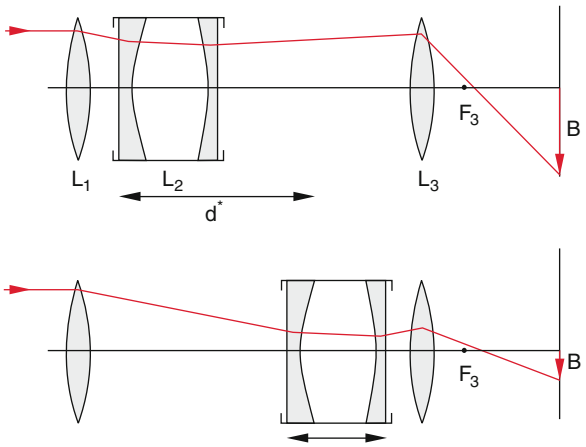


Fig. 9.36 Zoom-lens system. Alteration of the magnification M by shifting the pair of diverging lenses. L_2 within the distance d^* between the two converging lenses L_1–L_3

system is for an object distance $a \gg f$ proportional to the focal length of the system. In order to change the magnification by the factor $V = M_{max}/M_{min}$ one has to change the focal length by this factor. This can be achieved by shifting the pair of divergent lenses. The image plane should not change while doing this.

With a rather elaborate calculation [2], which starts from equations analogue to (9.33a–9.33c) one can show that the shift d at a given focal length f is related to the magnification ratio V by

$$d^* = \frac{(V - 1)}{\sqrt{V}}$$

Nowadays Zoom-objectives for cameras can be bought, which change their magnification by a factor 10 just by turning the lens mount thus converting the camera operation from wide angle to telephoto (Fig. 9.37) [3].

Example

$V = 4 \rightarrow d^* = 1.5 \cdot f.$

9.5.6 Lens Aberrations

The previous representation of lenses and their imaging properties were dealing with ideal lenses and are for real lenses approximately correct only for paraxial light rays close to the symmetry axis.

For rays which are farther away from the symmetry axis or which are inclined against this axis image aberrations occur, which cause that an object point A is no longer imaged into an image point B but rather into an area around B. This results in a blur of the image and often also in a distortion which is different for the different local parts of the

Fig. 9.37 Zoom-objective Zoomar MG 189

image. These deviations from an ideal image formation are called *aberrations.*

For all applications where very small structures should be still imaged true to scale all lens aberrations should be made as small as possible. Examples are the lithographic production of integrated circuits where a spatial resolution of 100 nm is required. Also the imaging of finer details in biological cells demands a distortion free image. Therefore large efforts are undertaken to realize systems of lenses with minimum aberrations.

We will now shortly discuss the most important lens aberrations and measures to minimize them [4–6].

9.5.6.1 Chromatic Aberration

Since the refractive index $n(\lambda)$ depends on the wavelength λ the focal length f changes with λ according to (9.25a). For instance, for glass $n(\lambda)$ increases with decreasing wavelength (normal dispersion, see Sect. 8.2). For incident parallel white light the focal point for blue light is therefore located before that for red light (Fig. 9.38). This can be demonstrated for a large audience when concentric circular rings

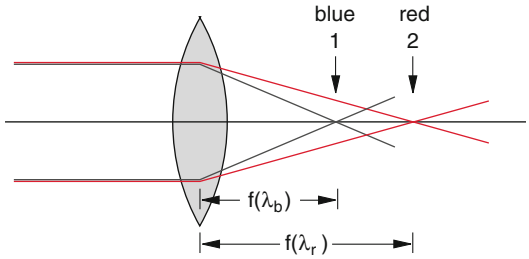


Fig. 9.38 Chromatic aberration

carved into a black coated plate are illuminated by a continuous light source and imaged onto a screen. For the position 1 in Fig. 9.38 one observes blue rings with red edges whereas in the position 2 red rings with blue edges are seen.

The chromatic aberration can be minimized when a system of two lenses with different refractive indices $n_1(\lambda)$ and $n_2(\lambda)$ is used. Such an achromat (Fig. 9.39) consists of a biconvex converging lens L_1 $(n_1\lambda)$ and a diverging lens L_2 with $n_2(\lambda)$ which are cemented together.

We will now calculate the relation between the focal lengths f_1 and f_2 that results in a refractive index n of the system which is nearly independent of λ. The lens Eq. (9.25a) gives for the focal length of the lens L_i

$$\frac{1}{f_i} = (n_i - 1)\varrho_i, \tag{9.34a}$$

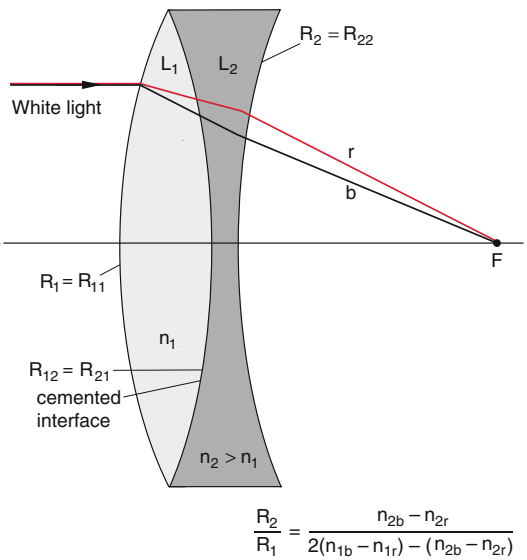


Fig. 9.39 Achromat with the two lenses cemented together

where $\varrho_i = (R_{i2} - R_{i1})/(R_{i2} \cdot R_{i1})$ and R_{i1}, R_{i2} are the radii of curvature for the front side and the backside of the lens L_i.

The focal length of the achromat with two lenses at close contact ($d = 0$) is

$$\frac{1}{f} = (n_1 - 1)\varrho_1 + (n_2 - 1)\varrho_2. \tag{9.34b}$$

The focal length of the achromat for blue light is equal to that of red light if

$$\begin{aligned} &(n_{1r} - 1)\varrho_1 + (n_{2r} - 1)\varrho_2 \\ &= (n_{1b} - 1)\varrho_1 + (n_{2b} - 1)\varrho_2 \\ &\Rightarrow \frac{\varrho_1}{\varrho_2} = -\frac{n_{2b} - n_{2r}}{n_{1b} - n_{1r}}. \end{aligned} \tag{9.34c}$$

The wavelength-dependence $n(\lambda)$ of the various types of glasses can be obtained from tables of the glass producers which give accurate values of $n(\lambda_i)$ for selected wavelengths λ_i that correspond to readily accessible spectral lines of some elements. Examples are:

$n_r = n(\lambda = 644$ nm = red cadmium line)
$n_D = n(\lambda = 590$ nm = mid between the two yellow sodium lines)
$n_g = n(\lambda = 546$ nm = green mercury line)
$n_b = n(\lambda = 480$ nm = blue cadmium line)
$n_h = n(\lambda = 404.65$ nm = violet mercury line)

For these wavelengths the refractive indices are listed for all current types of glasses.

Often the focal length of a lens is given for yellow light ($\lambda = 590$ nm) or for green light ($n_g = n(\lambda = 546$ nm)). The refractive index for green light is given as the average of the indices for red and blue light.

$$n_g \approx \frac{1}{2}(n_b + n_r).$$

From (9.34a) we obtain the ratio of the two focal lengths of the two lenses

$$\frac{f_2}{f_1} = \frac{(n_{1g} - 1)(n_{2b} - n_{2r})}{(n_{2g} - 1)(n_{1b} - n_{1r})}. \tag{9.34d}$$

With the abbreviation

$$v = \frac{n_g - 1}{n_b - n_r} = \frac{\delta_g}{\Delta\delta} \tag{9.34e}$$

introduced by Ernst Abbe one obtains

$$\frac{1}{f_1 v_1} = \frac{1}{f_2 v_2}. \tag{9.34f}$$

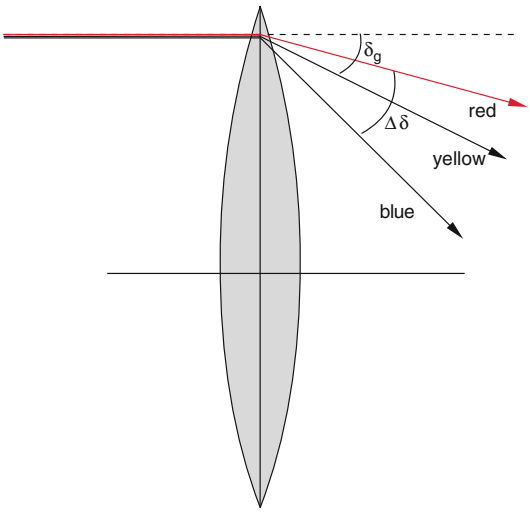


Fig. 9.40 Definition of the Abbe-number

The Abbe number v gives the ratio of diffraction angle δ_g for yellow light to the difference $\Delta\delta$ of the refraction angels for blue and red light. It is therefore a measure for the dispersion of a glass (Fig. 9.40).

The total focal length of the system with two lenses can be calculated with the equation

$$\begin{aligned} \frac{1}{f} &= \frac{1}{f_1} + \frac{1}{f_2} = -\frac{v_1}{v_2}\frac{1}{f_2} + \frac{1}{f_2} \\ &= \frac{v_1 - v_2}{v_2}\frac{1}{f_2}. \end{aligned} \tag{9.34g}$$

With (9.34f) this yields for the focal lengths of the two lenses

$$f_1 = f \cdot \frac{v_1 - v_2}{v_1}; \quad f_2 = -f \cdot \frac{v_1 - v_2}{v_2}. \tag{9.34h}$$

In case of two lenses cemented together the two radii of curvature $R_{12} = R_{21}$ must be equal. If the first lens is symmetrical biconvex, it is $R_{11} = R_1 = -R_{12} = -R_{21}$ and $R_{22} = R_2$. This gives the condition

$$\begin{aligned} R_1 &= \frac{2(n_{1g} - 1)((v_1 - v_2))}{v_1}; \\ R_2 &= \frac{2(v_1 - v_2) \cdot f}{\frac{2v_2}{n_{2g}-1} - \frac{v_1}{n_{1g}-1}}. \end{aligned} \tag{9.34i}$$

Example

If the lens L_1 is made of optical glass BK7 ($n_{1b} = 1.52283$; $n_{1g} = 1.5168$; $n_{1r} = 1.51472$) and the lens L_2 of flint glass ($n_{2b} = 1.77647$; $n_{2g} = 1.75513$; $n_{2r} = 1.74843$ we obtain from (9.34e)

$v_1 = 64$; $v_2 = 27$.

If the focal length of the achromat shall be $f = 100$ mm, we get from (9.34h) the focal lengths of the two lenses: $f_1 = 64$ mm; $f_2 = -179$ mm => $R_1 = 66$ mm; $R_2 = +130$ mm. The diverging lens is therefore not symmetric.

9.5.6.2 Spherical Aberration

Also for monochromatic light one observes deviations from the correct point to point imaging. For example, the focal length of a lens with spherical surfaces depends on the distance h of the rays from the symmetry axis (Fig. 9.41). This ***spherical aberration***, which we have already discussed in Sect. 9.3 for the spherical mirror, occurs for thin as well as for thick lenses.

We will illustrate this at first for the refraction at a spherical surface (Fig. 9.42). When a paraxial ray with the distance h from the symmetry axis is incident onto a spherical surface the focal length is

$$f = R + b \text{ with } b = R \cdot \frac{\sin\beta}{\sin\gamma}.$$

With $\sin\beta = \sin\alpha/n$, $\sin\alpha = h/R$ and $\alpha = \beta + \gamma$ we get

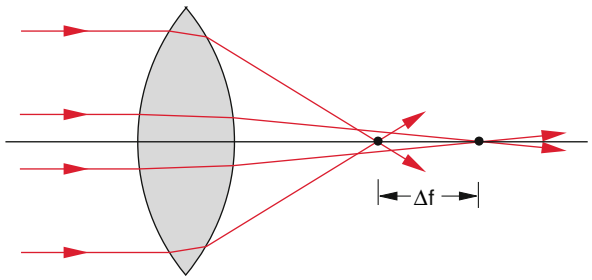


Fig. 9.41 Spherical aberration at the imaging by a spherical biconvex lens

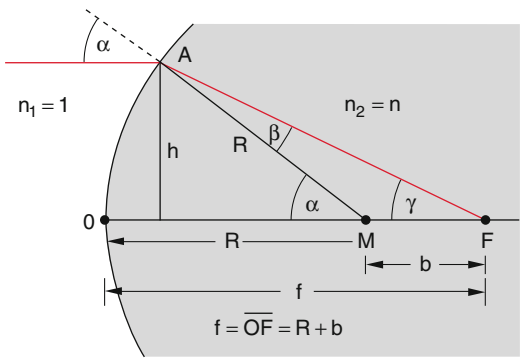


Fig. 9.42 Illustration of the dependence $f(h)$ of the focal length f from the distance h of the parallel ray from the symmetry axis

$$\begin{aligned} f &= R + \frac{h}{n \cdot \sin\gamma} = R\left[1 + \frac{1}{n\cos\beta - \cos\alpha}\right], \\ &= R \cdot \left[1 + \frac{1}{\sqrt{n^2 - \sin^2\alpha} - \sqrt{1 - \sin^2\alpha}}\right], \\ &= R \cdot \left[1 + \frac{1}{n \cdot \sqrt{1 - \frac{h^2}{n^2R^2}} - \sqrt{1 - \frac{h^2}{R^2}}}\right]. \end{aligned} \tag{9.35}$$

Neglecting the term $h^2/R^2 \ll 1$ completely, we obtain immediately the approximation (9.21b) for paraxial rays with small distances h. For a better approximation we expand the square root in (9.35) according to $\sqrt{(1-x)} = 1 - 1/2x$ for $x \ll 1$. We get with $(1-x)^{-1/2} \approx 1 + x$ after a short calculation the result

$$\begin{aligned} f &= R \cdot \left[\frac{n}{n-1} - \frac{h^2}{2n(n-1)R^2}\right] \\ &= f_0 - \Delta f(h). \end{aligned} \tag{9.36}$$

with $\Delta f(h) = R \cdot h^2/(2n(n-1)R^2)$

This shows that the focal length f decreases with increasing distance h of the rays from the axis.

In a similar way one obtains the imaging equation for a refracting spherical surface when the term h^2/R^2 is included (this is identical with the approximation $\cos\gamma \approx 1 - \gamma^2/2$).

$$\frac{1}{a} + \frac{n}{b} = \frac{n-1}{R} + h^2\left[\frac{1}{2a}\left(\frac{1}{a} + \frac{1}{R}\right)^2 + \frac{n}{2b}\left(\frac{1}{R} - \frac{1}{b}\right)^2\right], \tag{9.37}$$

which reduces to (9.22a–9.22c) for $(h \to 0)$.

The second term in (9.37) describes the deviation of the image distance b due to spherical aberration. This term depends on h and R as well as on the object distance a.

The focal length of a thin lens, taking into account spherical aberration can be obtain in a similar way as Eq. (9.25a) without aberration. One has to replace (9.21a) and (9.21b) by the more accurate Eq. (9.36) and instead of the paraxial approximation $\sin\alpha \approx \tan\alpha \approx \alpha$ the expansion

$$\sin\alpha \approx \alpha - \frac{1}{3!}\alpha^3 \text{ and } \cos\alpha \approx 1 - \frac{\alpha^2}{2}$$

must be used. The somewhat lengthy calculation yields for the difference of the reciprocal focal length with spherical aberration from the calculation without aberration

$$\Delta_s = \frac{1}{f(h)} - \frac{1}{f(h=0)}$$

when inserting the expressions (9.35) for $f(h)$ and (9.25a) for f_0

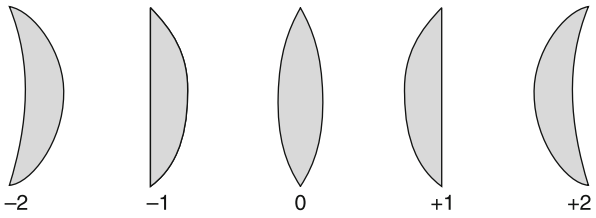


Fig. 9.43 Values of $q = (R_1 + R_2)/(R_1 - R_2)$ for some types of lenses

$$\Delta_s = \frac{h^2}{8f_0^3 n(n-1)^2}\left[n^3 + (3n+2)(n-1)^2 p^2 + 4(n^2-1)pq + (n+2)q^2\right] \tag{9.38}$$

with $q = (R_1 + R_2)/(R_2 - R_1)$ and $p = (b-a)/(b+a)$, where a and b are the object distance and the image distance.

The minimum spherical aberration is obtained for

$$q = -\frac{2(n^2-1)p}{n+2}. \tag{9.39}$$

In Fig. 9.43 the values of q are indicated for several types of lenses.

This illustrates that for the imaging of a far distant object ($a = \infty \rightarrow p = -1$) by a plan-convex lens with $n = 1.5$ and with the curved surface towards the object (9.39) the optimum value $q = 0.7$ is achieved. This is close to the value $q = 1$ for the aberration-free lens.

For a given refractive index n there is an optimum form of the lens for which Δ_s becomes minimum. For instance is it better for a plan-convex lens to turn the curved surface towards the object (Fig. 9.44b) because then the rays far away from the symmetry axis pass the lens around the minimum deviation (see Sect. 9.4).

The general rule for achieving minimum spherical aberration for the imaging by a plan-convex lens with object distance a and image distance b is the following:

Fort $a > b$ the curved surface should be directed towards the object, for $a < b$ it should be directed towards the image space. For a biconvex lens with $R_1 \neq R_2$ the side with the smaller value of R (more strongly curved) should be directed towards the object.

The spherical aberration can be decreased

- when the rays far from the axis are suppressed by an aperture
- when using a plan-convex lens where the convex surface is directed towards the object
- by the combination of several collecting and diverging lenses which realizes a spherical corrected lens system
- by special lenses with non-spherical surfaces. They are more difficult to ground but with modern computerized techniques it is nowadays possible to fabricate such lenses with nearly zero spherical aberrations. Since it takes more efforts to ground non-spherical lenses often such lenses are made of plexiglas, (acryl glass) which can be cast from its liquid phase in the wanted form and then needs only some polishing.

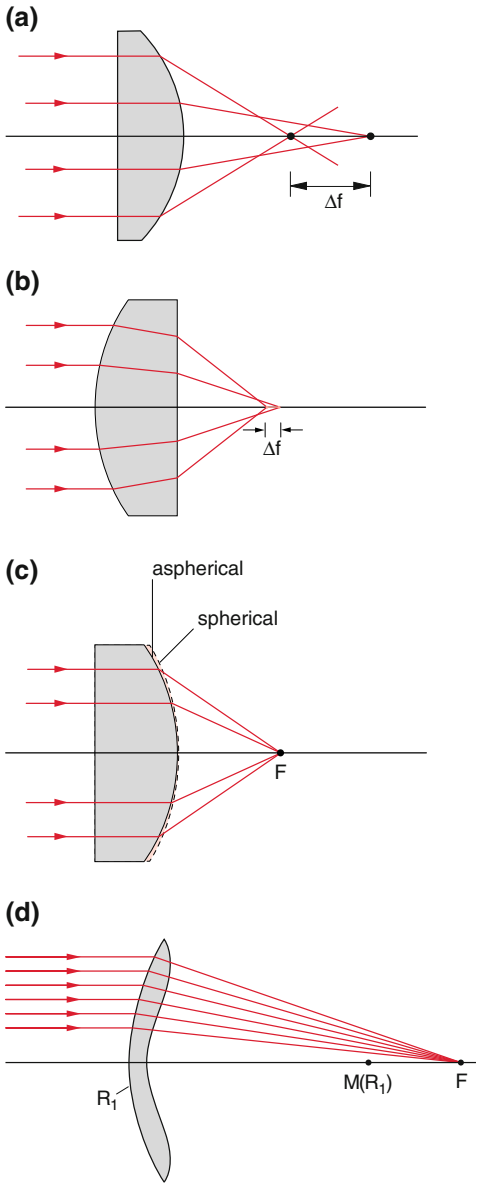


Fig. 9.44 Different spherical aberration by a plane-convex lens for different orientations of the lens. **a**) Plane surface towards the object **b**) spherical surface towards the object **c**) plane-convex aspherical lens **d**) convex-concave aspherical lens

9.5.6.3 Aspherical Lenses

An aspherical lens has at least one surface that deviates from a sphere or a plane. In Fig. 9.45 the difference is shown between a convex-plane lens with spherical front surface and an aspherical lens with a front surface deviating from a

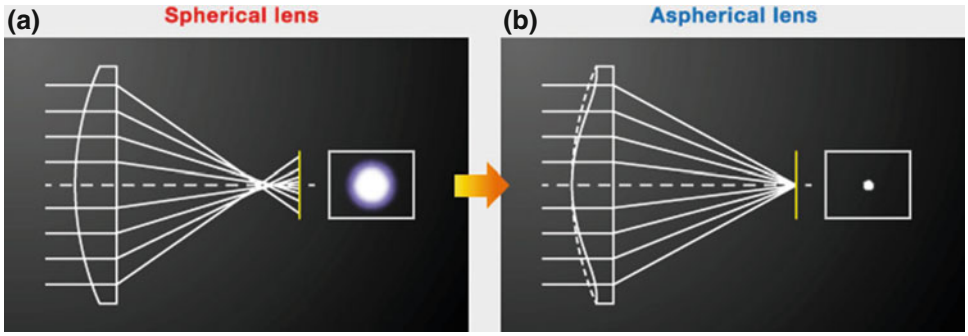


Fig. 9.45 Comparison of a plane-convex lens with spherical aberration **a**) and an aspherical lens where the spherical aberration is completely corrected **b**) [Edmund Optics]

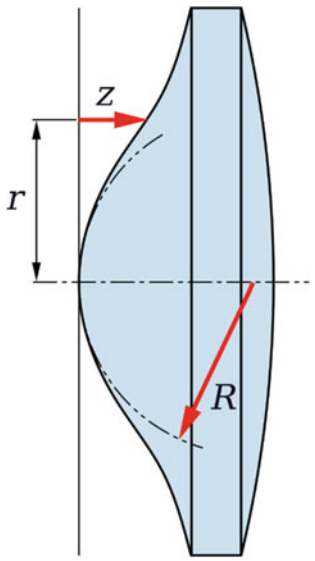


Fig. 9.46 Aberration-free imaging by an aspherical lens with a plane backside and an aspherical frontside. The radius of curvature changes with the distance h from the symmetry axis [Wikipedia]

sphere and a plane backside. In former times the problem of producing aspherical lenses was the lack of suitable polishing procedures which could make the wanted aspherical surface with the required accuracy (surface roughness smaller than $\lambda/10$). Nowadays computer-aided grounding and polishing machines are available, that can manufacture any desired surface form.

The radius of curvature of an aspheric surface changes with the distance h from the symmetry axis. This is indicated in Fig. 9.46 by the thin spherical dashed curves.

It is much cheaper to produce aspherical lenses made of plastic. They can be manufactured in special aspheric molds where the liquid plastic is poured in and pressed into the wanted form (Fig. 9.47). The problem is that the solidified lens shrinks against the liquid form. This has to be taken into account if a minimum accuracy of the lens form is required. Because the plastic material is softer than glass the polishing is easier.

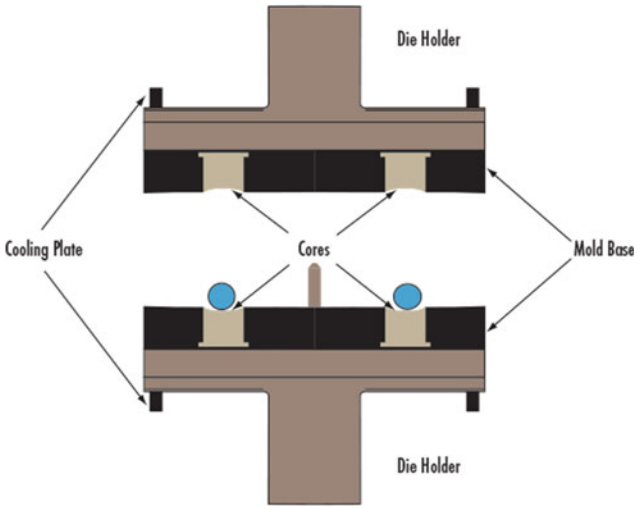


Fig. 9.47 Fabrication of photo-polymer aspherical lenses [Edmund Optics]

9.5.6.4 Coma

The spherical aberration has been discussed in Sect. 9.5.6.2 for incident rays parallel to the symmetry axis. If a parallel light beam passes through a lens that is inclined against the beam axis, (Fig. 9.48a) the refraction angles not only depend on the distance h of light rays from the symmetry axis, but they differ also for equal values of h above and below the axis. The focal points of the different light bundles no longer are located on the symmetry axis, which we choose as the x-axis.

When an object point A is imaged by the inclined lens the image points $B_i(x, y)$ for the different light bundles are at different values of x and y (Fig. 9.48b). If only spherical aberration would be present, the image of A would be a small circle around the image point B in the plane $x = x_B$. Because of the coma aberration the image of A is now a complex surface non-uniformly illuminated.

The effect is particularly obvious, when the central part of the lens is masked so that only rays through the outer parts of the lens contribute to the image. One obtains then instead

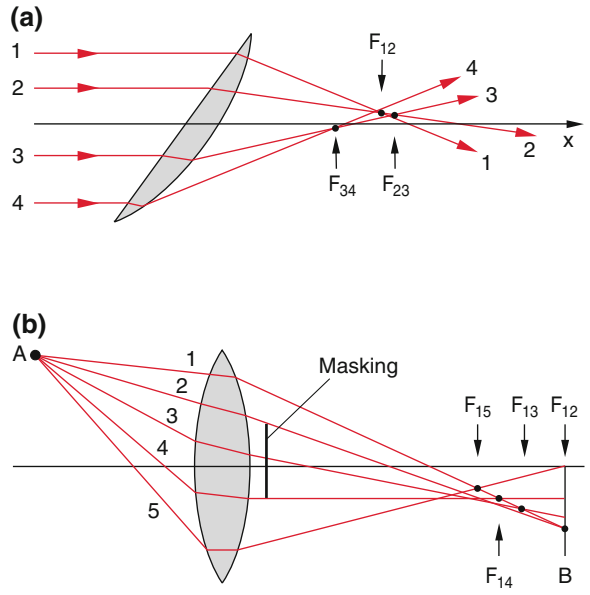


Fig. 9.48 **a**) Coma occurring when a light beam passes an inclined lens. The different rays have different focal points. **b**) For the imaging of a point A outside the symmetry axis the different rays have different images B away from the symmetry axis

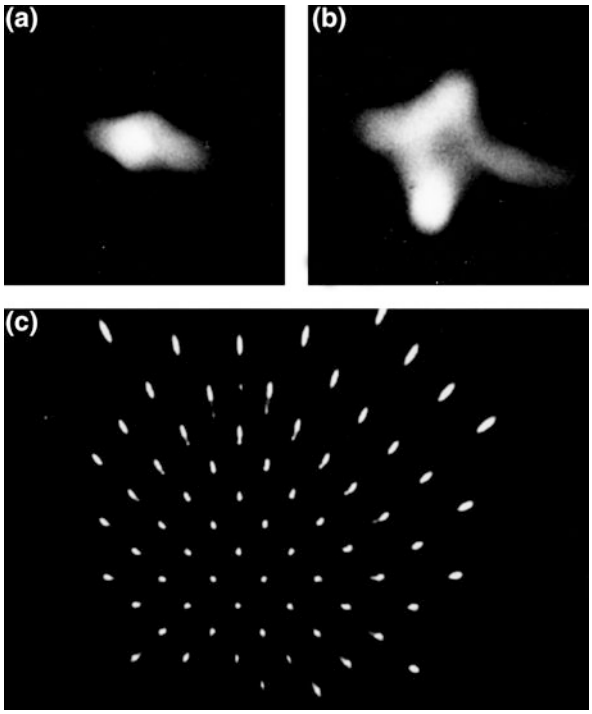


Fig. 9.49 Distorted images of a point A obtained with the arrangement of Fig. 9.47b. **a**) without mask, **b**) masking the central part of the lens **c**) distorted image o f a sheet with equally distributed holes

of a single image point B a blurred curve with a form that depends on the distance x_B of the image plane from the lens. In Fig. 9.49 such blurred images are shown for illustration of the effect. They have been obtained with the arrangement of Fig. 9.48b at the image plane $x = F_{13}$ with and without masking the central part of the lens.

This image distortion is called coma (greek: κομη = hair)

Remark Comets have their name from their striking tail which was believed by the Greek to be the hair of a goddess.

9.5.6.5 Astigmatism

Imaging of object points A far away from the symmetry axis, which is often necessary in photographic practice, results in a further aberration, called *astigmatism.* We will shortly discuss the distortion of images because it also appears often at the imaging of objects by our eye lens. We regard in Fig. 9.50a a horizontal plane S (*sagittal plane*) and a vertical sectional plane M (*meridian plane*) through an inclined light bundle, which is emitted by an object point A away from the symmetry axis and imaged by the lens into the image space.

All rays in the sagittal plane S are imaged into the image point B_S at the distance x_S from the lens whereas the rays in the meridian plane are imaged into another point B_M at the distance x_M. This is due to the fact, that the ray AM_1 hits the lens surface under a larger angle than the ray AM_2 and are therefore more strongly refracted.

The imaging of the object point A by all rays produces a horizontal line in the image plane b_M and a vertical line in the plane b_s, while in between these planes an elliptical image is produced (Fig. 9.50b)

The distance $\Delta x = b_S - b_M$ (astigmatic difference) becomes larger with increasing inclination of the incident light rays.

Such an astigmatic distortion also appears when a light beam passes through an oblique plane parallel glass plate (Fig. 9.51). When an oblique glass plate is placed into the optical path behind the imaging lens L the image of an object point A on the symmetry axis is no longer a point B but a horizontal or vertical line or an elliptical area around the image point B without distortion, depending on the image distance x_B.

Particularly pronounced are astigmatic aberrations for imaging by cylindrical lenses (Fig. 9.52), which focusses only in one direction. All rays from the object point A in a plane perpendicular to the cylinder axis (red plane in Fig. 9.52) are focused into the point b in this plane. All rays in a plane parallel to the cylinder axis are divergent. Their rearward prolongation intersect in the virtual image point B'. Overall is the image of the object point A a line parallel to the cylinder axis.

The correction of astigmatic aberrations can be achieved with a combination of cylindrical and spherical lenses. This can be for instance realized by a single lens, when a spherical lens gets an additional cylindrical curvature. This is used in eyeglasses for the correction of astigmatic aberrations of the eye.

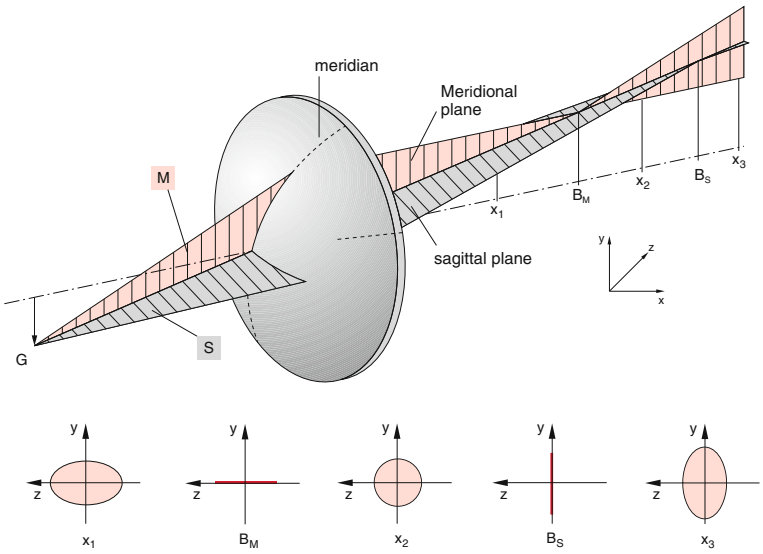


Fig. 9.50 Astigmatism at the imaging of an inclined light beam. **a**) Perspective view **b**) cross section of the light beam at different distances behind the lens

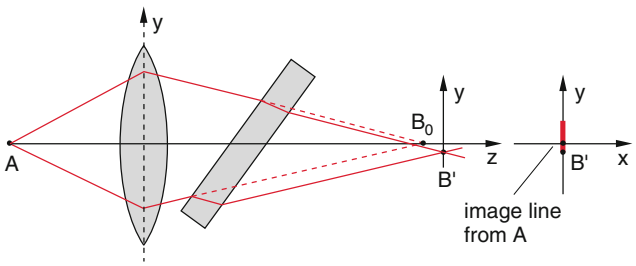


Fig. 9.51 Astigmatism of a light beam passing through an inclined plane-parallel plate behind the focusing lens. Without plate the image of A would be at B_0. The light rays in the horizontal cut (x-y-plane) intersect along the line B', they therefor do not form a point-like image

9.5.6.6 Image Field Curvature and Distortion

Due to the different refraction of light rays which enter a lens under different angles against the symmetry axis the image distances b_i are different for object points A_i in the same object plane but with different distances from the symmetry axis. The image of the object plane is no longer a flat plane but a curved surface (Fig. 9.53). Because of the astigmatic aberrations two different image distances x_S and x_M occur for the sagittal and the meridian rays. The image surfaces of the object plane are therefore two curved surfaces B_S and B_M. They can be visualized by rotating the image points B_i of object points A_i in a plane perpendicular to the symmetry axis around this axis.

This image field curvature can be demonstrated by imaging of a plane spoke wheel through a cylindrical lens (Fig. 9.54). Depending on the distance x_B of the image plane either the inner rings are focused/(Fig. 9.54a) or the inner rings (Fig. 9.54b).

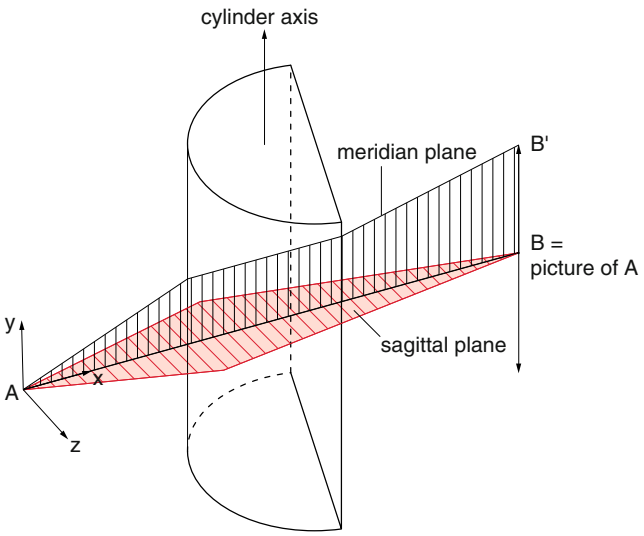


Fig. 9.52 Astigmatism at the imaging by a cylindrical lens

We had seen in Sect. 9.5.6.2, that masking the outer parts of a lens diminishes for paraxial ray the spherical aberration and improves the quality of the image. However, for rays inclined against the symmetry axis aberrations occur in spite of masking the edge rays, which results in the distortion of the image of extensive objects. This can be demonstrated by imaging a plane quadratic grid (Fig. 9.55). Placing a circular aperture in front of the lens which transmits only the central rays, the image shows a barrel distortion (Fig. 9.55 right),

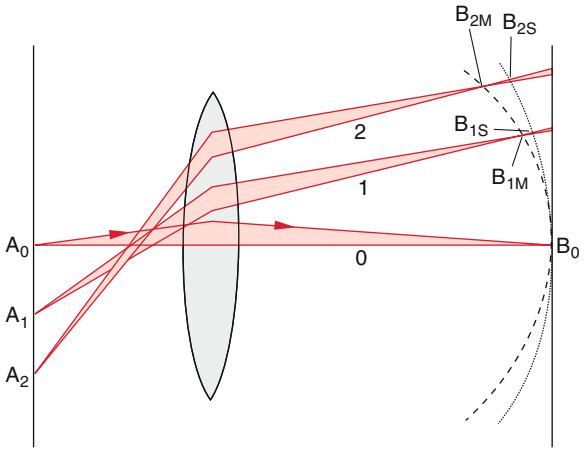


Fig. 9.53 Image field curvature

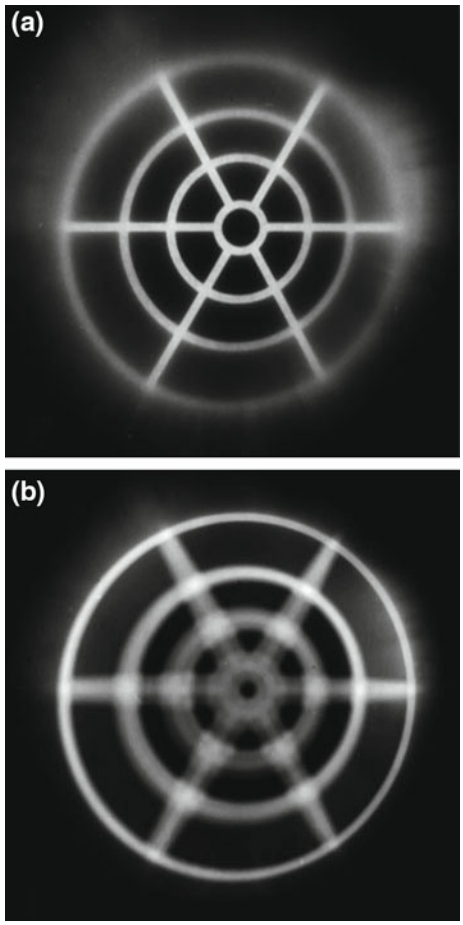


Fig. 9.54 Experimental demonstration of image field curvature of a plane spoke wheel. **a**) Image plane is at B_0 in Fig. 9.55 **b**) image plane is shifted towards the lens and includes B_{1M}

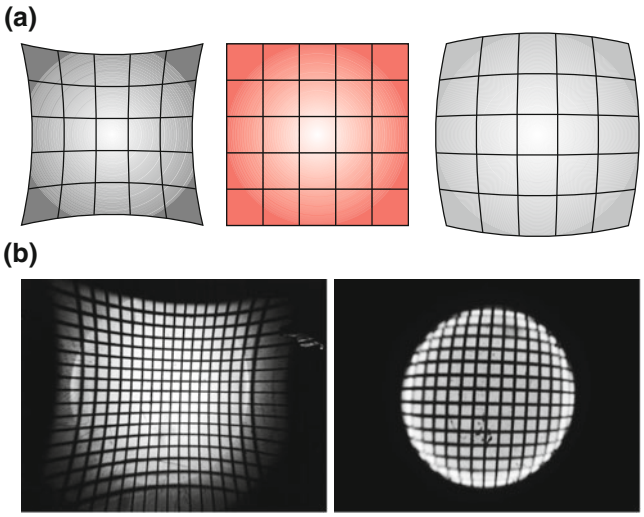


Fig. 9.55 Pin-cushion distortion and barrel distortion of a plane regular grid. **a**) Schematic representation (https://de.wikipedia.org/wiki/Abbildungsfehler) **b**) real photos, taken by the author

whereas a mask before the lens which blocks the central rays a pincushion distortion of the grid appears (Fig. 9.55 left).

This can be understood as follows:

We regard in Fig. 9.53 two points A_0 and A_1 of the extended plane object. Because of the larger refraction of the inclined rays from A_1 the image plane of B_1 is formed before that of B_0. Therefore the object point A_1 forms in the image plane of B_0 a circle with the center M, where M is defined by the dashed line through the center of the lens. The diameter $D = R_1R_2$ of this circle is determined by the edge rays from A_1 which can still pass through the aperture before the lens (Fig. 9.56a). Only rays within a narrow angular range around the ray 1 can be transmitted. A square around A_1 is imaged into a barrel like distorted area around M.

In Fig. 9.56b the circular aperture is placed behind the lens. Now again the point A_1 is imaged into a circular area with the center M_1 by rays in a narrow angular range around the ray 1. These rays form in the image plane of B_0 again a circle but with smaller diameter $D = R_1R_2$ and a center M_1 that is farther away from B_0 than M. Since the shift of M against M_1 becomes larger with increasing distance h of A_1 from the symmetry axis a square around A_1 is imaged into a pin-cushion distortion in the image plane of B_0.

9.5.7 The Aplanatic Imaging

For practical applications not only single points but extended areas should be imaged with minimum distortion. Furthermore a large luminosity of the imaging lens system is

required which excludes small apertures which restrict the transmitted light to paraxial rays close to the symmetry axis. Great efforts have been undertaken to construct lens systems which minimize all lens aberrations even for large aperture ratios. An important insight was provided by a relation, postulated by *Ernst Abbe* (1840–1905) between the magnification ratio $M = |B|/|A| = b/a$ of a lens system and the aperture angles u_g and u_b of the transmitted rays on the object side and the image side (Fig. 9.57). This relation states that even for large aperture angles an image formation with small distortions is possible, if **Abbe's sinus condition**

$$\frac{\sin u_g}{\sin u_g} = \frac{|B|}{|A|} = \text{const.} \tag{9.40}$$

is fulfilled.

We will illustrate Abbe's sinus condition by a simple example: The imaging of an illuminated circular aperture by a lens (Fig. 9.57). We assume that the aperture is illuminated by an extended light source with diameter A at a far distance. The parallel light bundles emitted by two different points of the source are drawn in Fig. 9.57 together with their phase planes. The path difference between upper and lower edge of the aperture is

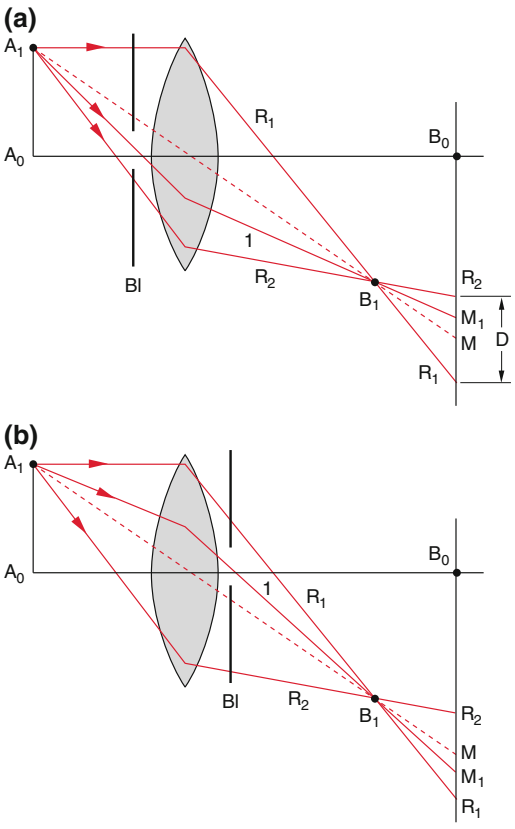


Fig. 9.56 For the imaging of a plane object the kind of distortion depends on the position of the aperture **a**) before and **b**) behind the lens

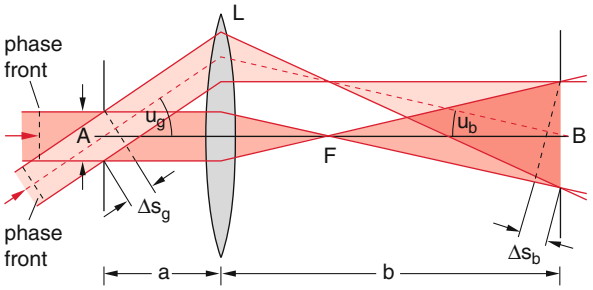


Fig. 9.57 Abbe's-sine-condition for the aplanatic imaging of an aperture area

$$\Delta s_g = A \cdot \sin u_g.$$

The lens images the aperture into the image plane at the distance b from the lens. The image of the aperture has the diameter B. For $B \ll b$ the curvature of the phase front can be neglected. The corresponding optical path difference on the image side is then

$$\Delta s_g = B \cdot \sin u_b.$$

For a distortion-free image the path difference Δs_g on the object side must be equal to Δs_b on the image side, because then every point of A is imaged into the plane of B. This gives immediately the sine-condition (9.40).

The image obeying the sine- condition is called **aplanatic** [7].

A single lens or a lens system can fulfill the conditions for an aplanatic imaging only for certain ranges Δa in the object space and Δb in the image space, which depend on the construction of the optical system [2, 9].

Figure 9.58 shows two examples of Zeiss photo-objectives which minimize lens aberrations. They are corrected for spherical and chromatic aberrations and also for astigmatism. The chromatic aberration is minimized by using an achromat in the *Tessar* (4 lenses) and a double achromat in the *Planar* (6 lenses). These objectives provide

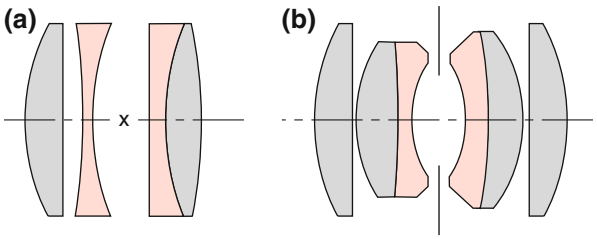


Fig. 9.58 The photo-objectives **a**) *Tessar* with data collected in Problem (9.14) **b**) *Planar*. Both objectives have been developed by Carl Zeiss, Jena. They provide an achromatic imaging and a broad realization of an aplanatic imaging

a nearly aberration-free imaging up to an opening ratio of 1:2.8 for a short focal length of the *Planar*. They were one of the standard objectives for better cameras (Fig. 9.58).

9.6 Matrix Methods of Geometrical Optics

The propagation of light rays through complex optical systems consisting of several lenses is generally complicated and not easy to calculate. Therefore methods have been developed which can perform such calculations for general optical systems with computers more readily and fast. A very efficient procedure is the matrix method, which will be shortly introduced.

In geometrical optics every optical imaging is described by the graphical depiction of light rays (see Sect. 9.1). These rays propagate in homogeneous media on straight lines which change their direction at the interfaces between two media with different refractive index n.

In optical systems with a symmetry axis each point $P(x, r)$ on the light ray is defined by its coordinate x on the symmetry axis and its distance $r = (y^2 + z^2)^{1/2}$ from this axis. The light rays can also be inclined against the symmetry axis by the angle α.

We can therefore describe the propagation of a light ray even through complicated systems by defining for each point of the ray its distance r from the axis and the inclination angel α of the ray in this point.

9.6.1 The Translation Matrix

Within the paraxial approximation (the distance r of a ray is small compared with the relevant distances in the x-direction, for instance the focal length of a lens) we can use the approximation $\sin\alpha \approx \tan\alpha \approx \alpha$. For the straight propagation of a light ray in a homogeneous medium from the plane $x = x_0$ to the plane $x = x_1$ the linear relations holds

$$\begin{aligned} r_1 &= (x_1 - x_0)\cdot\alpha_0 + r_0, \\ n\cdot\alpha_1 &= n\cdot\alpha_0. \end{aligned} \tag{9.41}$$

These linear equations for r and α can be written in the form of matrices. We describe the ray parameters r and α by a two-component column vector. We can then write (9.41) in matrix form as

$$\begin{pmatrix} r_1 \\ n\cdot\alpha_1 \end{pmatrix} = \begin{pmatrix} 1 & \frac{x_1 - x_0}{n} \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} r_0 \\ n\cdot\alpha_0 \end{pmatrix}. \tag{9.41a}$$

Denoting the distance between the planes $x = x_1$ and $x = x_0$ by $d = x_1 - x_0$ the translation matrix for a light ray in a homogeneous medium with refractive index n can be written as

$$\tilde{T} = \begin{pmatrix} 1 & d/n \\ 0 & 1 \end{pmatrix}. \tag{9.41b}$$

Note The angle α is defined as positive, if one proceeds from the x-axis anti-clockwise to the ray and as negative for the clockwise rotation.

9.6.2 The Refraction Matrix

Also for the refraction at the interface between two different media a linear relation exists between the parameters (r_1, α_1) on one side of the interface and $(r_2 = r_1, \alpha_2)$ on the other side. Using Snell's refraction law for small angles

$$n_1\alpha = n_2\beta$$

we obtain from Fig. 9.59 for a curved boundary with curvature radius R the relation

$$\alpha - \alpha_1 = -\alpha_2 + \beta = \gamma = r_1/R,$$

(α_2 is negative, because it is measured clockwise against the positive x-axis.

$$\begin{aligned} & n_1\left(\frac{r_1}{R} + \alpha_1\right) = n_2\left(\frac{r_1}{R} + \alpha_2\right) \\ & \Rightarrow n_2\alpha_2 = n_1\alpha_1 + (n_1 - n_2)\frac{r_1}{R}. \end{aligned}$$

We therefore obtain the equations for the refraction at a spherical surface (Fig. 9.59)

$$\begin{aligned} r_2 &= r_1, \\ n_2\alpha_2 &= n_1\alpha_1 + (n_1 - n_2)r_1/R, \end{aligned} \tag{9.42a}$$

We get from (9.42a) the matrix equation

$$\begin{pmatrix} r_2 \\ n_2\alpha_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \frac{n_1 - n_2}{R} & 1 \end{pmatrix} \cdot \begin{pmatrix} r_1 \\ n_1\alpha_1 \end{pmatrix}. \tag{9.42b}$$

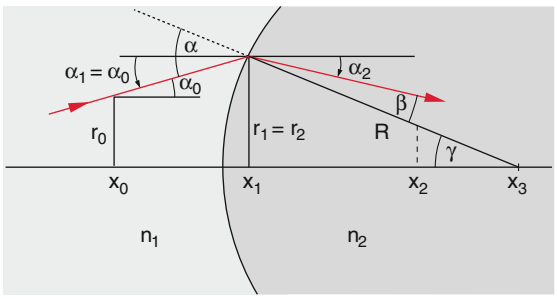


Fig. 9.59 Matrix representation of translation and refraction of a light ray

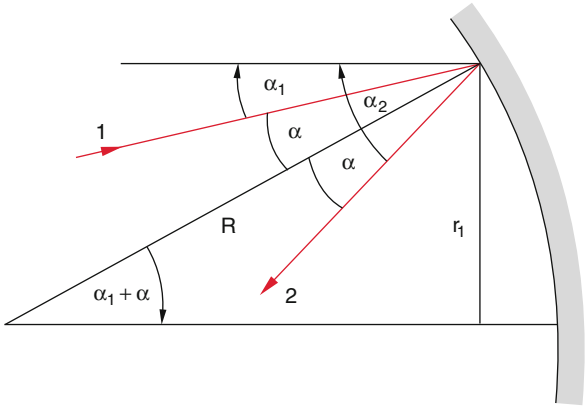


Fig. 9.60 Characteristic quantities for the matrix representation of a spherical mirror

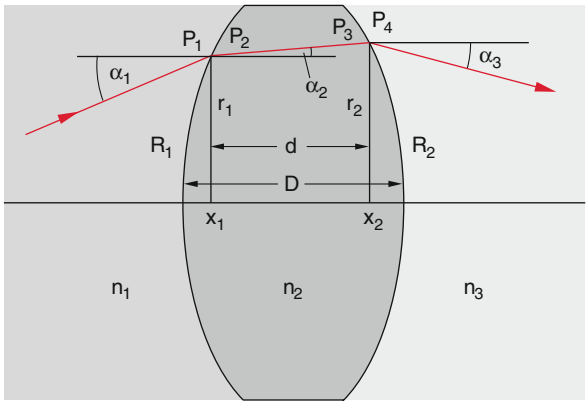


Fig. 9.61 Illustration for the transformation matrix of a lens with radii of curvature R_1 and R_2 and thickness D

The refraction at a curved surface with radius R between two media with refractive indices n_1 and n_2 can be therefore described by the refraction matrix

$$\tilde{B} = \begin{pmatrix} 1 & 0 \\ \frac{n_1 - n_2}{R} & 1 \end{pmatrix} \tag{9.42c}$$

9.6.3 Reflection Matrix

Analogue to the refraction at a spherical interface we can describe the reflection at a spherical mirror by the reflection matrix. From Fig. 9.60 we get with a mirror radius R and the refractive index $n = 1$ in the space left of the mirror the relations

$$\begin{aligned} r_2 &= r_1 \\ \alpha_2 &= 2\alpha + \alpha_1 = 2(\alpha + \alpha_1) - \alpha_1 \\ &= -2\frac{r_1}{R} - \alpha_1 \end{aligned}$$

when the direction of the angle arrow in Fig. 9.59 and the note in the previous section, regarding the sign of the angle is taken into account

$$\tilde{R} = \begin{pmatrix} 1 & 0 \\ -\frac{2}{R} & -1 \end{pmatrix}, \tag{9.43}$$

The matrix method allows the calculation of the path of light rays through optical systems with many refracting or reflecting interfaces by multiplication of the corresponding matrices of the different interfaces. This technique is particular useful for the calculation of complex lens systems such as camera objectives or microscopes. We will illustrate this by some examples.

9.6.4 Transformation Matrix of a Lens

We regard in Fig. 9.61 a light ray, which passes from the object space with refractive index ($n = n_1$) through a lens with thickness D, radii of curvature R_1 and R_2 and refractive index n_2 into the image space with $n = n_3$. The ray parameters (n, r, α) change at every interface from the initial values (n_1, r_1, α_1) at the point P_1 to the final values (n_3, r_3, α_3) in the point P_4. The sequence of the parameters can be written as

$$\begin{pmatrix} r_1 \\ n_1\alpha_1 \end{pmatrix} \rightarrow \begin{pmatrix} r_1 \\ n_2\alpha_2 \end{pmatrix} \rightarrow \begin{pmatrix} r_2 \\ n_2\alpha_2 \end{pmatrix} \rightarrow \begin{pmatrix} r_2 \\ n_3\alpha_3 \end{pmatrix}.$$

In the matrix notation initial and final state are related by

$$\begin{pmatrix} r_2 \\ n_3\alpha_3 \end{pmatrix} = \tilde{B}_2 \cdot \tilde{T}_{12} \cdot \tilde{B}_1 \begin{pmatrix} r_1 \\ n_1\alpha_1 \end{pmatrix} \tag{9.44}$$

with the matrices

$$\tilde{B}_1 = \begin{pmatrix} \frac{n_1 - n_2}{R_1} & 0 \\ 1 & 1 \end{pmatrix}; \tag{9.44a}$$

$$\tilde{T}_{12} = \begin{pmatrix} 1 & \frac{x_2 - x_1}{n_2} \\ 0 & 1 \end{pmatrix}; \tag{9.44b}$$

$$\tilde{B}_2 = \begin{pmatrix} -\frac{n_2 - n_3}{R_2} & 0 \\ 1 & 1 \end{pmatrix}; \tag{9.44c}$$

where, according to the rules of matrix multiplication at first the input vector (n_1, α_1, r_1) is multiplied by the matrix B_1 and the resulting vector (n_2, α_2, r_2) is multiplied with T_{12} etc.

The radius of curvature R_2 of the second lens surface is negative, according to the definition in Sect. 9.5.2 whereas R_1 is positive.

The product of the three matrices is the transformation matrix M_L of an arbitrary lens with refractive index n_2 in a surrounding with refractive indices n_1 and n_3. Performing the multiplication gives

$$\tilde{M}_L = \tilde{B}_2\tilde{T}_{12}\tilde{B}_1 = \begin{pmatrix} 1-\frac{x_{21}n_{21}}{n_2R_1} & \frac{x_{21}}{n_2} \\ \frac{n_2n_{32}R_1 - n_2n_{21}R_2 - n_{32}n_{21}x_{21}}{n_2R_1R_2} & 1+\frac{n_{32}x_1}{n_2R_2} \end{pmatrix}, \tag{9.45}$$

For thin lenses ($x_2 - x_1 => 0$) with the focal length f and the refractive index $n_2 = n$ in air ($n_1 = n_3 = 1$) the equation (9.45) has the much simpler form

$$\tilde{M}_L = \begin{pmatrix} 1 & 0 \\ (n-1)\left(\frac{1}{R_2}-\frac{1}{R_1}\right) & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 \\ -\frac{1}{f} & 1 \end{pmatrix}, \tag{9.45a}$$

where we have used the relation (9.25a) for the focal length f.

9.6.5 Imaging Matrix

When the object point A is imaged by the lens L onto the image point B (Fig. 9.62) the imaging equation in matrix representation is with $n_1 = n_3 = n_1, n_2 = n$

$$\begin{pmatrix} r_2 \\ \alpha_2 \end{pmatrix} = \tilde{M}_{AB}\begin{pmatrix} r_1 \\ \alpha_1 \end{pmatrix} \tag{9.46}$$

with the imaging matrix

$$\tilde{M}_{AB} = \tilde{T}_2\tilde{M}_L\tilde{T}_1, \tag{9.47}$$

where the translation matrices for the object space and the image space are

$$\tilde{T}_1 = \begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}; \quad \tilde{T}_2 = \begin{pmatrix} 1 & b \\ 0 & 1 \end{pmatrix}, \tag{9.48}$$

The imaging matrix for thin lenses is then, using (9.45a)

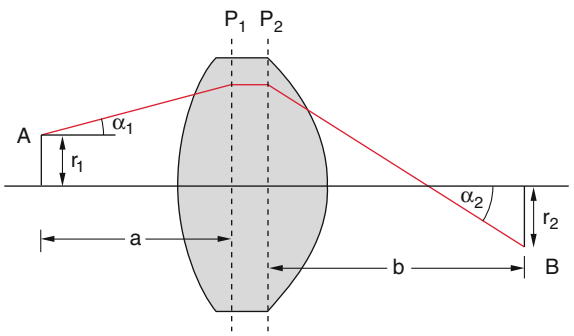


Fig. 9.62 Illustration of the transformation matrix of a thick lens

$$\tilde{M}_{AB} = \begin{pmatrix} 1-\frac{b}{f} & a+b-\frac{ab}{f} \\ -\frac{1}{f} & 1-\frac{a}{f} \end{pmatrix}. \tag{9.49}$$

The imaging equation is therefore

$$\begin{pmatrix} r_2 \\ \alpha_2 \end{pmatrix}_B = \begin{pmatrix} \left(1-\frac{b}{f}\right)r_1 + \left(a+b-\frac{ab}{f}\right)\alpha_1 \\ -\frac{r_1}{f} + \left(1-\frac{a}{f}\right)\alpha_1 \end{pmatrix}, \tag{9.46a}$$

where a and b are the distances from the medium plane of the thin lens to object a or image b (Fig. 9.26). For thick lenses the distances are measured up to the principal planes (Fig. 9.30).

For $\alpha_1 = 0$ (incident rays are parallel to the symmetry axis) is

$$\begin{pmatrix} r_2 \\ \alpha_2 \end{pmatrix} = \begin{pmatrix} \frac{f-b}{f}\cdot r_1 \\ -\frac{r_1}{f} \end{pmatrix}, \tag{9.46b}$$

Such rays intersect the symmetry axis in the image space behind the lens at $r_2 = 0 => f = b$. The image of an infinitely far away object is formed in the focal point of the lens.

9.6.6 Matrices of Lens Systems

The advantage of the matrix method becomes more evident when systems of many lenses shall be calculated. We will this illustrate for the example of a system of two lenses with the focal lengths f_1 and f_2, the distances d_{ik} between the principal planes of each lens and the distances D between the two lenses (Fig. 9.63).

The transformation matrix of the lens system is

$$\tilde{M}_{LS} = \tilde{B}_4\tilde{T}_{34}\tilde{B}_3\tilde{T}_{23}\tilde{B}_2\tilde{T}_{12}\tilde{B}_1 \tilde{M}_{L2}\tilde{T}_{23}\tilde{M}_{L1}, \tag{9.50}$$

where $\tilde{T}_{ik}$ is the transformation matrix for the light path from P_i to P_k and $\tilde{B}_i$ the matrix (9.42c) for the refraction at the surface i with radius of curvature R_i. According to (9.45). As already indicated in (9.50) the outer thre matrices can be condensed into the matrix M_{Li} of the thick lenses L_1 and L_2.

For further examples see the problems (9.12–9.14) and the references [1, 2, 8, 9].

9.6.7 Jones Vectors

As has been mentioned already in the beginning of this chapter, the polarization characteristics of light can be formally treated within the frame work of geometrical optics if we introduce the electric field vector $\boldsymbol{E}$ in addition to the vector $\boldsymbol{k}$ which gives the direction of propagation of the light

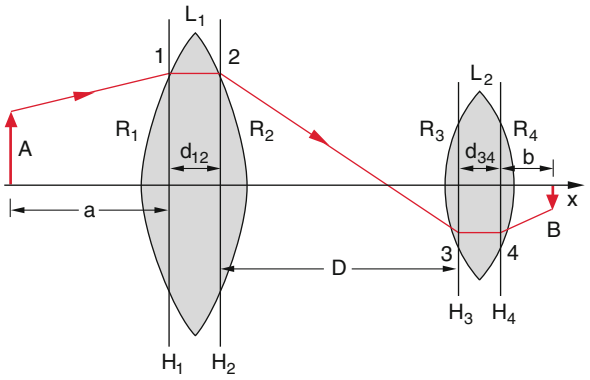


Fig. 9.63 Characteristic quantities needed for the transformation matrix for the imaging by a lens system of two lenses

ray. When we choose the z-axis as the direction of propagation the electric field vector becomes

$$\boldsymbol{E} = E_x\hat{e}_x + E_y\hat{e}_y,$$

where the components E_x and E_y may be complex numbers (see Sect. 7.4). We therefore write the polarization vector as column vector

$$\boldsymbol{E} = \begin{pmatrix} E_x \\ E_y \end{pmatrix} = \begin{pmatrix} E_{0x}e^{i\varphi_x} \\ E_{0y}e^{i\varphi_y} \end{pmatrix} \tag{9.51a}$$

With $|E| = \sqrt{E_x^2 + E_y^2}$ we can define a normalized vector

$$\boldsymbol{J} = \begin{pmatrix} J_x \\ J_y \end{pmatrix} = \frac{1}{|E|}\begin{pmatrix} E_{0x}e^{i\varphi_x} \\ E_{0y}e^{i\varphi_y} \end{pmatrix} \tag{9.51b}$$

called the Jones vector. Since the polarization characteristics of the electric vector only depends on the phase difference $\Delta\varphi = \varphi_x - \varphi_y$ between the two components but not on the absolute value of the phases, we can choose $\varphi_x = 0$.

Examples

(a) For light linearly polarized in x-direction is $E_{0y} = 0$ and the Jones vector becomes

$$\boldsymbol{J}_\mathrm{h} = \begin{pmatrix} 1 \\ 0 \end{pmatrix} \quad \text{(horizontal Polarization)}$$

when we choose the x-direction as the horizontal and the y-direction as the vertical direction.
For light polarized in the y-direction we get correspondingly

$$\boldsymbol{J}_\mathrm{v} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} \quad \text{(vertical Polarization)}.$$

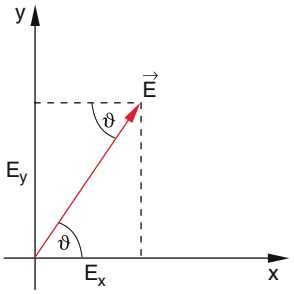


Fig. 9.64 Electric vector $\boldsymbol{E}$ of linear polarized light

(b) When the E-vector points into the direction ϑ against the x-axis (Fig. 9.64) we have $E_x = E \cdot \cos\vartheta$ and $E_y = E \cdot \sin\vartheta$. The two phases φ_x and φ_y are equal and we can choose them as $\varphi_x = \varphi_y = 0$. The Jones vector then becomes

$$\boldsymbol{J}(\vartheta) = \begin{pmatrix} \cos\vartheta \\ \sin\vartheta \end{pmatrix}, \tag{9.52}$$

This reduces for $\vartheta = 45°$ to

$$\boldsymbol{J}_{45°} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{9.52a}$$

(c) For circular polarized light is $E_{0x} = E_{0y} = E/\sqrt{2}$ and $\varphi_x - \varphi_y = \pm\pi/2$. For σ^+-light the Jones vector becomes

$$J(\sigma^+) = \frac{1}{E}\begin{pmatrix} E_{0x} \\ E_{0x} \cdot e^{i\pi/2} \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ +i \end{pmatrix} \tag{9.53a}$$

while for σ^--light we get correspondingly

$$J(\sigma^-) = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -i \end{pmatrix}. \tag{9.53b}$$

For elliptical polarized light is

$$E = \begin{pmatrix} E_x \\ E_y e^{-i\varphi} \end{pmatrix},$$

The Jones vector therefore becomes

$$\boldsymbol{J} = \frac{1}{|E|}\begin{pmatrix} E_x \\ E_y e^{-i\varphi} \end{pmatrix}$$

If polarized light propagates through anisotropic media or if it is reflected at inclined surfaces its polarization characteristics changes (see Chap. 8). Such polarization changing elements can be described analogue to lenses by matrices, called *Jones matrices*. For instance, a linear

polarizer, which transmits light with its E-vector in x-direction, is described by

$$M_{(x)} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} (x\text{−polarizer}) \tag{9.54a}$$

When unpolarized light passes through an x-polarizer the transmitted light has the E-vector

$$E_{(t)} = \begin{pmatrix} E_{tx} \\ E_{ty} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} E_{ex} \\ E_{ey} \end{pmatrix} = \begin{pmatrix} E_{ex} \\ 0 \end{pmatrix} \tag{9.54b}$$

which has only a component in x-direction. A linear polarizer with a maximum transmission direction ϑ against the x-axis has the Jones matrix

$$M_{(\theta)} = \begin{pmatrix} \cos^2\theta & \sin\theta\cos^2\theta \\ \sin\theta \cdot \cos^2\theta & \sin^2\theta \end{pmatrix}. \tag{9.54c}$$

Examples

The Jones matrix for $\vartheta = 45°$ is

$$M_{(45°)} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}. \tag{9.54d}$$

An optical retardation plate causes a rotation of the polarization plane. It has the Jones matrix

$$M = \begin{pmatrix} e^{i\Delta\varphi x} & 0 \\ 0 & e^{i\Delta\varphi y} \end{pmatrix}, \tag{9.55}$$

The exit light is then

$$\begin{aligned} \begin{pmatrix} E_{tx} \\ E_{ty} \end{pmatrix} &= \begin{pmatrix} e^{i\Delta\varphi_x} & 0 \\ 0 & e^{i\Delta\varphi_y} \end{pmatrix} \cdot \begin{pmatrix} E_{ex}e^{i\varphi_x} \\ E_{ey}e^{i\varphi_y} \end{pmatrix} \\ &= \begin{pmatrix} E_{ex}e^{i(\varphi_x + \Delta\varphi x)} \\ E_{ey}e^{i(\varphi_y + \Delta\varphi y)} \end{pmatrix} \end{aligned} \tag{9.56}$$

For a $\lambda/4$-wave plate with the fast axis in the x-direction is $\Delta\varphi_y - \Delta\varphi_x = \pi/2$. The Jones matrix for the $\lambda/4$-plate is then

$$M_{\lambda/4} = e^{-i\frac{\pi}{4}}\begin{pmatrix} 1 & 0 \\ 0 & +i \end{pmatrix} = \frac{1}{2}\sqrt{2}\begin{pmatrix} 1-i & 0 \\ 0 & 1+i \end{pmatrix}, \tag{9.57}$$

where we have chosen $\Delta\varphi_x = -\pi/4$ and $\Delta\varphi_y = +\pi/4$.

For a $\lambda/2$-plate with the fast axis in x-direction we obtain the Jones matrix

$$M^{(x)}_{\lambda/2} = e^{-i\pi/2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} -i & 0 \\ 0 & +i \end{pmatrix}, \tag{9.57a}$$

while for the fast axis in y-direction we get

$$M^{(y)}_{\lambda/2} = e^{-i\pi/2}\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} +i & 0 \\ 0 & -i \end{pmatrix}. \tag{9.57b}$$

The polarization condition of light after passing through several elements that change the polarization status can be calculated by multiplying the corresponding matrices.

Example

A linear polarized wave with its $\boldsymbol{E}$-vector in the direction of 45° against the x-axis passes through a $\lambda/2$-plate with the fast axis in x-direction. Subsequently it traverses a $\lambda/4$-plate with the fast axis in y-direction. The exit wave is described by

$$\begin{aligned} \begin{pmatrix} E_{tx} \\ V_{ty} \end{pmatrix} &= \frac{E_0}{2}\sqrt{2}\begin{pmatrix} 1 & 0 \\ 0 & i \end{pmatrix}\begin{pmatrix} -i & 0 \\ 0 & +i \end{pmatrix} \\ &\quad \times \frac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \frac{E_0}{2}\begin{pmatrix} -i \\ -1 \end{pmatrix} = \frac{E_0}{2}e^{i\pi/2}\begin{pmatrix} 1 \\ -i \end{pmatrix}. \end{aligned}$$

This is a circular polarized σ^--wave with a phase that is shifted against the incident wave by $\Delta\varphi = \pi/2$. For further information see [10].

9.7 Geometrical Optics of the Atmosphere

Several optical phenomena in our earth atmosphere, which are related to reflection and refraction of sun light can be explained by geometrical optics. However, there are also phenomena such as light scattering (see Sect. 10.9) which can be only described by the wave model of light. Often quantum theory is needed to obtain a quantitative description e.g. of absorption and scattering of radiation in the atmosphere. The optics of our atmosphere is therefore much more complex than the few examples given here might suggest [11, 12]

9.7.1 Deflection of Light Rays in the Atmosphere

Since the density of the atmosphere decreases with increasing heights h (see Vol. 1, Sect. 7.2), also the refractive index $n(h)$ decreases with increasing h. If a light ray enters the atmosphere from the outside (e.g. from a star) under the angle ζ against the vertical, the radial change of

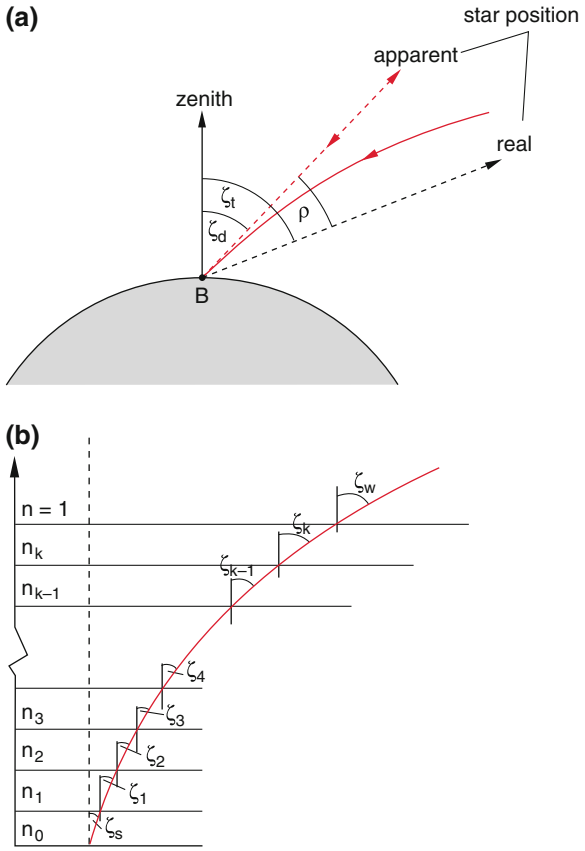


Fig. 9.65 Astronomical refraction of light in the atmosphere. The curvature of light from a star is here strongly exaggerated. **b)** Segmentation of the atmosphere into small sections

$n(h)$ cause a curvature of the ray (Fig. 9.65a). This is quantitatively illustrated in Fig: 9.65b.

This curvature of light rays in the atmosphere has the effect, that the angle ζ under which the light from a star appears for the observer (*zenith distance*) is decreased from the true value ζ_t to the apparent value ζ_a. The difference $\varrho = \zeta_t - \zeta_a$ between the true and the apparent zenith distance is called **refraction angle** of the atmosphere. It increases with the path length of the light ray through the atmosphere. It is therefore larger for stars which appear close to the horizon.

In order to determine the refraction angle ρ we divide the atmosphere into many thin horizontal layers (Fig. 9.65b). Within each of these layers the refractive index changes so little that we can regard it as constant. Within this approximation $n(h)$ makes a small jump from layer to layer and we approximate the continuous function $n(h)$ by a step function and the smooth path of the light ray by a polygon course.

We can apply Snell's law of refraction for each interface between successive layers and obtain with $n_0 = n(h = 0)$

$$\begin{aligned} n_0 \cdot \sin\zeta_s &= n_1 \cdot \sin\zeta_1 = n_2 \cdot \sin\zeta_2 = \ldots \\ &= n_k \cdot \sin\zeta_k = \sin\zeta_t \\ \Rightarrow \zeta_t &= n_0 \cdot \sin\zeta_a, \end{aligned} \tag{9.58}$$

because for large values of k the density of the atmosphere at large values of h becomes so low that $n_k = 1$.

The refraction angle ρ is very small. Therefore we can write in (9.58)

$$\begin{aligned} \sin\zeta_t &= \sin(\varrho + \zeta_a) = \sin\varrho \cdot \cos\zeta_a + \cos\varrho \cdot \sin\zeta_a \\ n_0 \cdot \sin\zeta_a &= \sin\zeta_t \approx \varrho \cdot \cos\zeta_a + \sin\zeta_a \\ \Rightarrow \varrho &= (n_0 - 1)\tan\zeta_a \end{aligned} \tag{9.59}$$

The experimental observation gives the value

$$\varrho_{exp} = 58.2'' \cdot \tan\zeta_a \quad \text{for} \quad \zeta < \zeta_{max} \quad \text{with } \zeta_{max} \approx 60^\circ$$

For $\zeta = 88.5^\circ$ is $\rho \approx 30'$. This corresponds nearly with the angular diameter of the sun (≈30′). This means: When the lower edge of the sun just touches the horizon the sun has in reality already sunk completely.

For the precise determination of star positions the refraction of the atmosphere has to be taken into account.

The refraction of the atmosphere has the effect that an observer at the heights h can see farther up to a point C (Fig. 9.66) than the straight line of the tangent from B to A would indicate. The point C appears to the observer B in the direction of C'. The horizon appears to be lifted by the angle $(\alpha_t - \alpha_a)$.

Example

At the heights h of the observer B the distance $\overline{AB} = \sqrt{(R+h)^2 - R_2} \approx \sqrt{2Rh}$. With $R = 6370$ km (earth radius) we obtain AB = 35.7 km. The refraction in the atmosphere increases this to BC = 38 km.

If the atmosphere shows inversion the temperature gradient dT/dh increases with h (instead decreasing under normal conditions) the gradient dn/dh becomes particularly large and with it the curvature of the light rays. One can then see "above" a

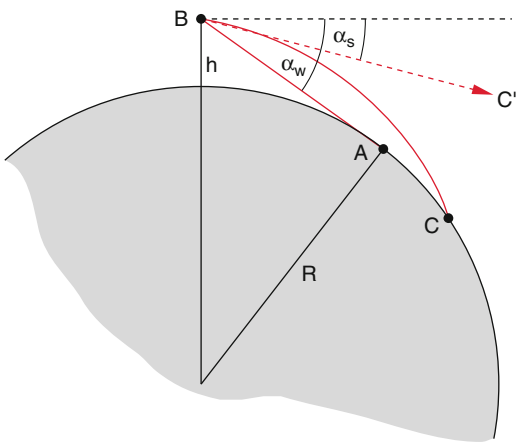


Fig. 9.66 Enlargement of sight distance due to refraction in the atmosphere

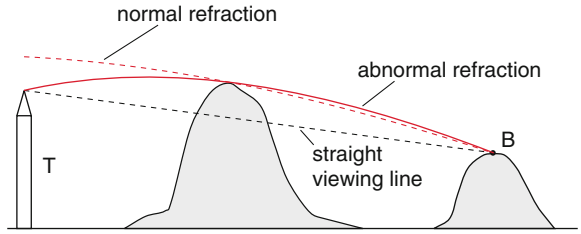


Fig. 9.67 Anomalous refraction of the atmosphere with $dT/dh > 0$, with resultant enlargement of the sight distance

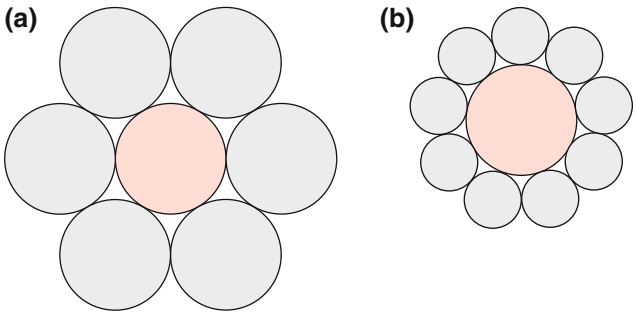


Fig. 9.68 Optical illusion. The circle in the center has the same size in **a**) as in **b**) although it appears larger in **b**

barrier (Fig. 9.67). For example the observer *B* in Fig. 9.67 can see the peak of a tower T which would be hidden behind the mountain for a light ray along a straight line.

9.7.2 Apparent Size of the Rising Moon

The rising full moon appears to the observer substantially larger than at its position high in the sky. This is often erroneously explained by the refraction of light in the atmosphere. The refraction plays only a minor role which let the full moon appear as a slightly elliptical disc just above the horizon. The apparent larger size of the moon close to the horizon is a pure psychological effect, an optical illusion. Our brain compares the moon diameter with its distance from the horizon. If the latter is small the diameter of the moon appears larger. This is illustrated by Fig. 9.68. The red circle in the mid has exactly the same size in the Figs. 9.68a and 9.68b. Nevertheless it appears larger in (b) than in (a), because it is surrounded in (b) by smaller circles, in (a) by larger ones.

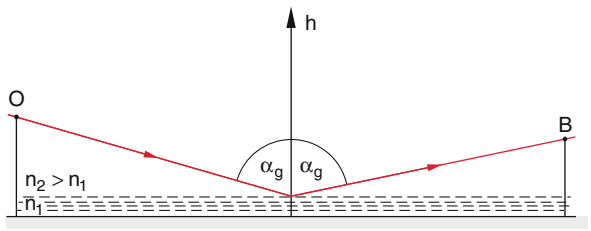


Fig. 9.69 Total reflection at an atmospheric layer close to the ground with a large gradient $dn/dh > 0$

If the moon is photographed at different positions all picture of the moon have the same size, because here the psychological effects is not present.

9.7.3 Fata Morgana

Also the Fata Morgana is based on the gradient $n(h)$ of the refractive index in the atmosphere causing reflection and curvature of light rays. The intense radiation of the sun in hot regions heats up the atmosphere close to the ground, in particular above surfaces that absorb much of the sun radiation (for instance black asphalt roads). In such cases a negative temperature gradient ($dT/dh < 0$) and a positive density gradient ($d\varrho/dh > 0$) occur. The gradient of the refractive index dn/dh then becomes especially large. Light rays which are incident from above in a nearly horizontal direction are totally reflected at the atmospheric layer closely above ground (Fig. 9.69). The observer *B* then sees through

Fig. 9.70 Fata morgana in the desert. The apparent lake is a reflection of the sky. The islands in the lake are in fact far away mountains

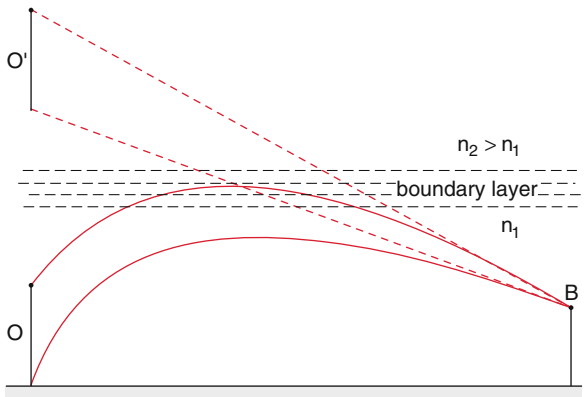


Fig. 9.71 Curvature of light rays in the atmosphere by sufficiently large gradient $dn/dh > 0$ of the refractive index n

the flickering atmosphere light from the blue sky, which appears on the road as a lake (Figs. 9.70).

During noon time the sand in the desert strongly heats up. Therefore such mirror images (Fat a Morgana) appear often in the desert and delude the thirsty walker lakes and plenty of water. Far mountains appear as islands in such simulated lakes.

Example

$T(h = 0) = 45\,°\text{C} = 318\,\text{K}$; $T(h = 50\ \text{m}) = 20\ °\text{C} = 293\ \text{K}$; $\varrho_0(T = 273\,\text{K}) = 1.293\,\text{kg/m}^3$; $\varrho_1(T = 318\,\text{K}) = 1.110\ \text{kg/m}^3$; $\varrho_2(T = 293\,\text{K}) = 1.205\,\text{kg/m}^3$. For the refractive index we obtain

$$n(\varrho) = n(\varrho_0) \cdot \frac{\varrho}{\varrho_0}$$

$$n_1 = 1 + 2.77 \times 10^{-4} \cdot \frac{1.110}{1.239} = 1.000238,$$

$$n_2 = 1.000277 \cdot \frac{1.205}{1.293} = 1.0002582.$$

The gradient of the refractive index is then $2 \times 10^{-5}/50\ \text{m} = 4 \times 10^{-7}/\text{m}$. The angle of total reflection is therefore

$$\sin \alpha_g = \frac{n_1}{n_2} = 0.999975 \Rightarrow \alpha_g = 89.59°.$$

Light rays which are incident under an angle $\alpha \leq \alpha_g$ are totally reflected.

When the temperature rises with increasing h the refractive index decreases with heights ($n_2(h) < n_1(h = 0)$). In this case a curvature of the rays entering from below onto the interface occurs (Fig. 9.71). Only for a sufficient large gradient d*n*/d*h* total reflection can be observed. An object O far away from the observer, appears for the observer B above its real location. If total reflection occurs the image of the object is reversed.

9.7.4 Rainbows

The colorful and impressive picture of a rainbow (Fig. 9.72) can be observed, when the sun, no longer obscured by clouds, shines onto a rain shower and the observers stands between sun and rain with his back towards the sun. The rainbow represents the visible segment of a circle with the center *M* on the extended line *SB* (Fig. 9.73). Only shortly before sunset a half circle can be seen.

Often two rainbows are observed. The primary rainbow has a sharp edge at its red outside. Towards the inside follow the colors with decreasing wavelength from red to blue. The opening angle between the symmetry axis *SBM* and the sharp red edge is $\varphi_H \approx 42°$ while the secondary rainbow with

Fig. 9.72 Primary Rainbow with weaker secondary bow. Note the reverse color sequence in the secondary bow

reversed color sequence has an opening angle of $\varphi_N = 51°$ between symmetry axis and the sharp blue outer edge.

Since the amplitudes of refracted and reflected light depend on the polarization of the light, the rainbow light is partially polarized.

Rene Descartes (1596–1650) recognized already in 1637 that the observed rainbow results from refraction and reflection of sunlight at many small water droplets in a rain front (Fig. 9.74). A light ray enters a water droplet, is refracted, passes through the droplet and is reflected at the backside (Fig. 9.75). At the exit it is again refracted. For the primary rainbow there occurs only one reflection (Fig. 9.75a), for the

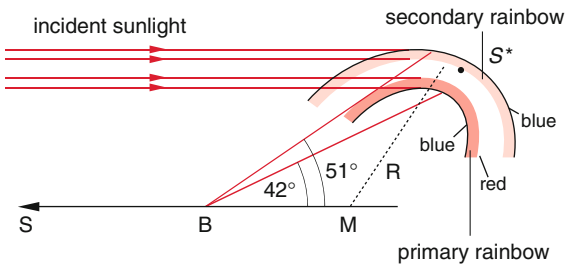


Fig. 9.73 Conditions for the observation of rain bows

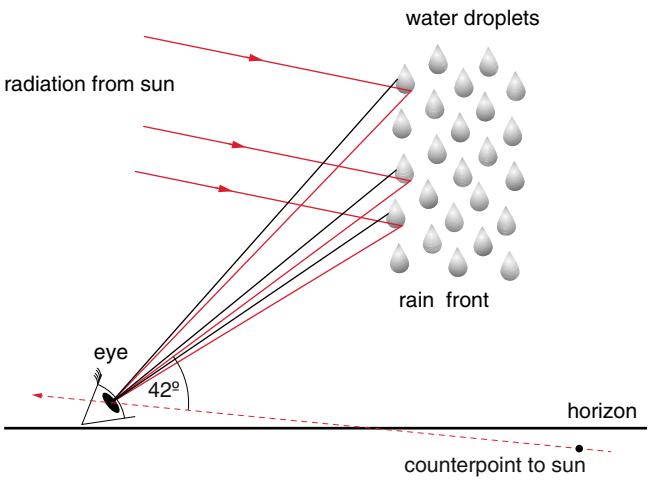


Fig. 9.74 Reflection and refraction of sunlight by many water droplets in a rain front

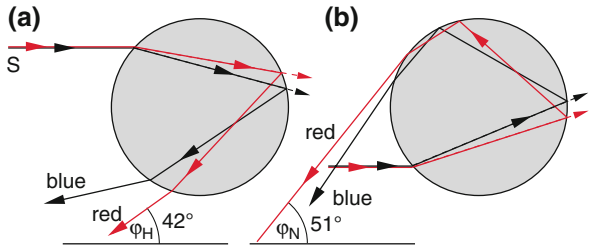


Fig. 9.75 Explanation of the generation of **a**) primary and **b**) secondary rain bow

secondary bow two reflections (Fig. 9.75b). This causes the reverse sequence of colors. Since the refractive index of water depends on the wavelength λ, the refraction angle differs for the different colors.

The deflection angle $\delta = 180° - \varphi$ for the light rays leaving the droplet depends on the entrance position z. From Fig. 9.76a we get the relations:

$$\delta = 180° - 4\beta + 2\alpha,$$
$$\sin\alpha = \frac{z}{R}; \sin\beta = \frac{z}{n \cdot R}.$$

The function $\delta(z)$ becomes minimum for (see Problem 9.10)

$$z = R \cdot \sqrt{\frac{1}{3}(4 - n^2)},$$

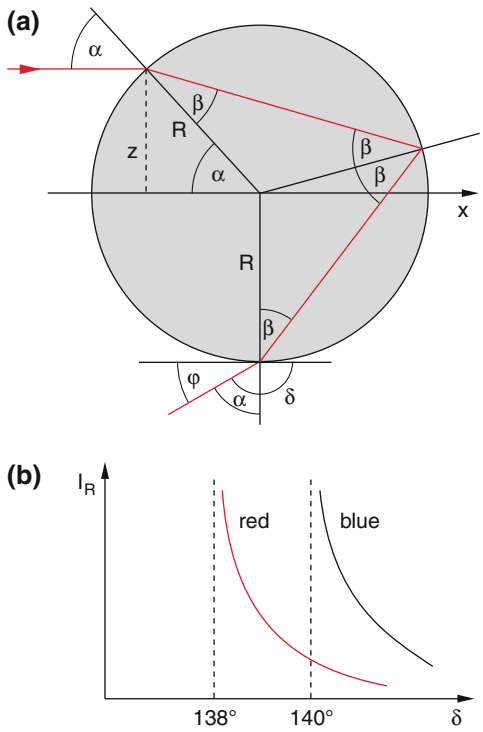


Fig. 9.76 **a**) Determination of the deflection angle δ as a function of the distance z of the incident sun light from the symmetry axis. **b**) Reflected intensity as a function of the deflection angle

For n = 1.33 the angle $\varphi = 4\beta - 2\alpha = 4\arcsin(z/nR)$ becomes $\varphi_{\max} = 42°$.

At this angle $\varphi_{\max}$ is $d\varphi/dz = 0$ and a maximum width Δz of incident rays contributes to the deflection of the sunlight into the angular interval $\varphi_{max} \pm \Delta\varphi$. Therefore for this angle the intensity of the sunlight reflected by the rain front becomes maximum. This is illustrated in Fig. 9.77. The light rays around the ray 6 (in the figure these are the rays 5–10) are all reflected aproximately into the same direction around $\varphi_H = 42°$.

For the two reflections leading to the secondary rainbow a similar consideration (see Problem 9.10) gives the angle $\varphi_N \approx 51°$.

The angular width $\Delta\varphi$ of the rainbow can be derived from the dispersion $n(\lambda)$ of water. The result is

$$\Delta\varphi = \left(\frac{d\varphi}{dn}\right) \cdot \left(\frac{dn}{d\lambda}\right) \cdot \Delta\lambda \text{ with } \Delta\lambda = \lambda_{\text{red}} - \lambda_{\text{blue}} \approx 330\,\text{nm}.$$

Although the theory of Descartes describes the main features of the rainbow correctly there are still finer details, such as additional faint red-green rainbows in between the primary rainbow, which cannot be understood by this theory. They are caused by interference and diffraction phenomena (see Chap. 10) and can be therefore only explained if the wave nature of light is considered [13].

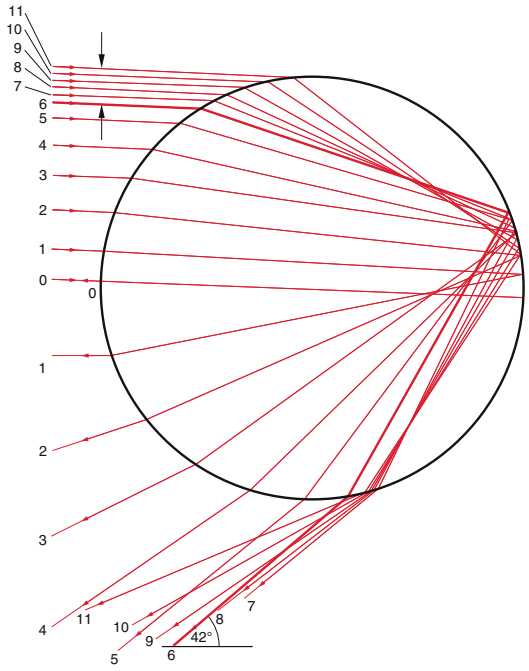


Fig. 9.77 Geometric construction of the rain bow effect, which appears as accumulation of rays for red light around the angle δ = 42° against the incident direction and for blue light around δ = 40°. The total deflection angle is then 180° − δ

Summary

- When diffraction phenomena can be neglected (this demands that the diameter of light beams is large compared to the wavelength λ) the propagation of light can be described by geometrical optics, which uses the concept of light rays.
- For the ideal optical imaging all rays emerging from a point A (light source) are imaged onto a point B (image of A). For real situations the image of A is an area around the image point B, The imaging can be realized by reflection (mirrors) or refraction (lenses).
- The imaging magnification is defined as the ratio of image diameter to object diameter
- The equation for imaging by thin lenses with focal length f is

$$\frac{1}{a}+\frac{1}{b}=\frac{1}{f},$$

  where a is the object distance AO from the lens and b the image distance OB.
- The total focal length f of a system of two close thin lenses L_1 and L_2 obeys the equation

$$\frac{1}{f_1}+\frac{1}{f_2}=\frac{1}{f}$$

- All optical elements (except the plane mirror) have imaging aberrations. They can be neglected for paraxial rays (their maximum distance from the symmetry axis is small compared to the focal length (paraxial approximation).
- The most important imaging aberrations are the spherical and the chromatic aberration, astigmatism, coma and image field curvature.
- The imaging by thick lenses can be reduced to that of thin lenses by introduction of principal planes.
- Systems of several lenses allow a broad variation of the imaging properties by changing the distance between the lenses without changing object and image planes.
- Within the paraxial approximation the imaging can be described by matrices. The imaging matrix of a system of lenses is the product of the matrices of the system components.
- In a medium with spatially varying refractive index $n(\boldsymbol{r})$ the light rays are curved. The curvature is proportional to grad $n(\boldsymbol{r})$.
- The phenomenon of the Fata Morgana appears if large gradients $\mathrm{d}n/\mathrm{d}h$ of the refractive index of air with the height h exist which cause a curvature of light rays
- Rainbows are caused by refraction and reflection of the sunlight in small water droplets.

Problems

9.1. Show by applying Fermat's principle, that a reflecting surface which focusses a plane wave into a point must be a paraboloid.

9.2. A plane wave is incident at the angle α onto a plane mirror. By which angle is the reflected wave turned, if the mirror is rotated by the angle δ? How is the situation for a spherical mirror, if the incident wave hits the mirror in the direction of the symmetry axis?

9.3. Derive Eq. (9.26) directly from Fig. 9.27 with the approximation $\sin x \approx \tan x \approx x$.

9.4. Between two plane mirrors at the positions $z \pm d/2$ is a point-like light source A placed at the position $z = 1/3d, x = 0$. Determine by graphical construction the four images B_i, which are closest to A.

9.5. A 2 cm thick water layer (n = 1.33) is located above a 4 cm thick layer of carbon tetrachloride (n = 1.46) in a cylindrical glass with radius R = 3 cm.
(a) What is the maximum angle of incidence α_{max} under which the center of the glass bottom can be still seen?
(b) How large must R be to allow $\alpha_{max} = 90°$?

9.6. You should produce with a lens a tenfold magnified image B of an object A at a distance d = 3 m from A. What is the focal length of the lens?

9.7. A light ray passes through a plane parallel glass plate with refractive index n and thickness d. The light ray enters the glass plate under the angel α.
(a) Show that the exit ray is parallel to the incident ray
(b) How large is the shift against the incident ray?

9.8. A light ray hits a mirror that consists of three plane surfaces which are orthogonal to each other. Show that the ray is reflected antiparallel to the incident ray, independent of the point of incidence.

9.9. A telescope has an objective lens with D_1 = 5 cm diameter and a focal length f_1 = 20 cm. How large should be the diameter D_2 of the ocular lens with focal length f_2 = 2 cm, to ensure that all light collected by the objective lens passes through the ocular lens? What is the angular magnification?

9.10. A light ray hits a glass sphere with radius R and refractive index n at a distance $z = h$ from the axis (Fig. 9.78). It is refracted at the first surface and reflected at the back surface.

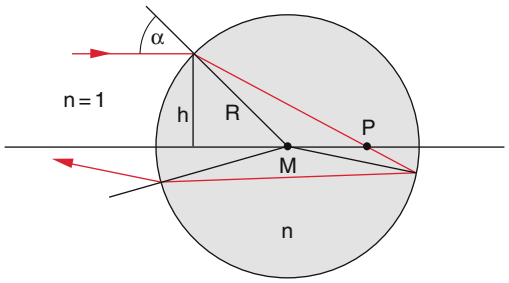


Fig. 9.78 To problem 9.10

(a) Where does the rays intersect the axis?
(b) Under which angle δ against the incident ray does it leave the sphere after one or two reflections?
(c) For which ratio h/R is δ minimum?
(d) Show that for water spheres (n = 1.33) $\delta_{min} = 138°$ for one reflection and 128° for two reflections.

9.11. A thin lens with R_1 = +10 cm and R_2 = +20 cm has the refractive index $n(\lambda = 600\ nm) = 1.485$ and $n(\lambda = 400\ nm) = 1.50$
(a) What are the focal lengths for the two wavelengths? (b) Give the parameters for a diverging lens that compensates the chromatic aberration.

9.12. Two thin lenses with f_1 and f_2 have a distance D ($D < f_1$ and $D < f_2$). What is the focal length of the lens system with f_1 = 10 cm, f_2 = 50 cm and D = 5 cm?

9.13. Two concave mirrors S_1 and S_2 with the centers of curvature M_1 and M_2 face each other at a distance d. Where are the images B_1 and B_2 of a point A located on the symmetry axis between S_1 and S_2 x cm away from S_1 generated by S_1 and S_2? Numerical example: R_1 = 24 cm; R_2 = 40 cm; d = 60 cm; x = 6 cm.

9.14. Calculate with the matric method the focal length of the special version of the Tessar lens system shown in Fig. 9.58a with the numerical data (given in cm)
$R_1 = 1.682$; $R_2 = -27.57$; $R_3 = -3.457$
$R_4 = 1.582$; $R_5 = \infty$; $R_6 = 1.92$; $R_7 = -2.40$
$n_1 = 1.6116$; $n_2 = 1.6053$; $n_3 = 1.5123$; $n_4 = 1.6116$;
$d_{12} = 0.357$; $d_{23} = 0.189$; $d_{34} = 0.081$;
$d_{45} = 0.325$; $d_{56} = 0.217$; $d_{67} = 0.396$.

References

1. Ch. Pendleburry: Lenses and System of Lenses, treatged after the manner of Gauss. (Forgotten Books 2018)
2. https://en.wikipedia.org/wiki/Zoom_lens
3. https://www.bing.com/search?q=camera+zoom+lenses&FORM=QSRE5
4. W.T. Welfored: Aberrations of optical systems (CRC-Press 1986)
5. https://en.wikipedia.org/wiki/Optical_aberration
6. C. Velzel: A Course in Lens design (Springer Series in Optical Sciences 183 (2014)
7. P. Zamora et.al. Aplanatic thin TIR Lens Proced. SPIE 8850, 101117 (2012)
8. A. Gerrard, J.M. Burch: Introduction to Matrix Methods in Optics Dover Books on Physics (Courier Corporation 1994)
9. Ironside: Matrix Methods in Optics, Cambridge Univ. Press 1995
10. Eugene Hecht: *Optics*. 4. Auflage. Addison-Wesley Longman, Amsterdam 2001
11. H. Stewart, R. Hopfield: Atmospheric Effects in: Applied Optics and Optical Engineering, ed. by R. Kingslake, Vol 1 (Academic Press New York 1965 p. 127–152
12. Earl J McCartney: Optics of the Atmosphere (Wiley 1976) Atmospheric Optics Wikipedia
13. Robert Greenler: Rainbows, Halos and Glories: Cambridge Univ. Press (1980)
14. Atmospheric Optics: Wkipedia with a detailled explanation and an extended reference list.

Interference, Diffraction and Scattering

10

Because of the linearity of the wave equation

$$\Delta \boldsymbol{E} = \frac{1}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} \tag{10.1}$$

every linear combination of arbitrary solutions $\boldsymbol{E}_1$ and $\boldsymbol{E}_2$

$$\boldsymbol{E} = a\boldsymbol{E}_1 + b\boldsymbol{E}_2$$

is also a solution of (10.1).

In order to obtain the total wave field $\boldsymbol{E}(\boldsymbol{r}, t)$ at an arbitrary point $P(\boldsymbol{r})$ at time t one has to add all amplitudes of the partial waves $\boldsymbol{E}_i(\boldsymbol{r}, t)$ that superimpose in P (superposition principle of linear equations). The total field amplitude

$$\boldsymbol{E}(\boldsymbol{r}, t) = \sum_m \boldsymbol{A}_m(\boldsymbol{r}, t)\mathrm{e}^{\mathrm{i}\varphi_m} \tag{10.2}$$

depends on the amplitudes $\boldsymbol{A}_m(\boldsymbol{r}, t)$ and on the phases φ_m of the superimposing partial waves. It is generally dependent on the local position $\boldsymbol{r}$ as well as on the time t.

This superposition of partial waves is called **interference** (see Vol. 1, Sect. 11.10). The total spatial area where the partial waves overlap is called the **interference field**. Its spatial structure is determined by the total intensity $I(\boldsymbol{r}, t) \propto |\boldsymbol{E}(\boldsymbol{r}, t)|^2$. Spatial limitations might suppress part of the interfering waves, which are then missing in the sum (10.2). This leads to incomplete interference which results in diffraction phenomena (see Sects. 10.7 and 11.3.4), causing additional structures of the wave field.

10.1 Temporal and Spatial Coherence

A temporal stationary interference structure can be only observed, if the phase differences $\Delta\varphi = \varphi_j - \varphi_k$ between arbitrary partial waves $\boldsymbol{E}_j$ and $\boldsymbol{E}_k$ in the point $P(\boldsymbol{r})$ do not change during the observation time Δt by more than 2π. The partial waves are then called **temporal coherent**. A possible time dependence $\Delta\varphi$ of the phase difference can have different reasons.

- The frequency ν may change in time
- The light source emits finite wave trains with randomly distributed phases.
- The refractive index of the medium between source and observer might fluctuate in time.

The maximum tine interval $\Delta t_{\max}$ during which the phase differences between all interfering partial waves are changing by less than 2π is the **coherence time**.

We will illustrate this by regarding a light source that emits light with the central frequency ν_0 and the spectral width $\Delta\nu$ (Fig. 10.1a). This light can be considered as the superposition of many partial waves with frequencies within the spectral range $\Delta\nu$.

The phase difference between these partial waves with frequencies $\nu_1 = \nu_0 - \Delta\nu/2$ and $\nu_2 = \nu_0 + \Delta\nu/2$ is

$$\Delta\varphi(t) = 2\pi(\nu_2 - \nu_1)(t - t_0).$$

It increases linearly with time t. After the coherence time $\Delta t_\mathrm{c} = 1/\Delta\nu$ it has increased to $\Delta\varphi(\Delta t_\mathrm{c}) > 2\pi$.

The coherence time of the light wave therefore equals the inverse spectral width $\Delta\nu$

$$\Delta t_\mathrm{c} = \frac{1}{\Delta\nu}. \tag{10.3}$$

For all other components with frequency differences $\Delta\nu < \Delta\nu_\mathrm{c}$ is $\Delta\varphi(\Delta t_\mathrm{c}) < 2\pi$. The superposition of all components contains all phase differences from 0 to 2π (Fig. 10.1b). For the temporal average of the superposition we get

$$\langle \boldsymbol{E}(\boldsymbol{r}, t)\rangle = \frac{1}{\Delta t_\mathrm{c}}\int_0^{\Delta t_\mathrm{c}} \sum_m \boldsymbol{A}_m(\boldsymbol{r})\mathrm{e}^{\mathrm{i}\varphi_m}\mathrm{d}t \equiv 0.$$

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_10

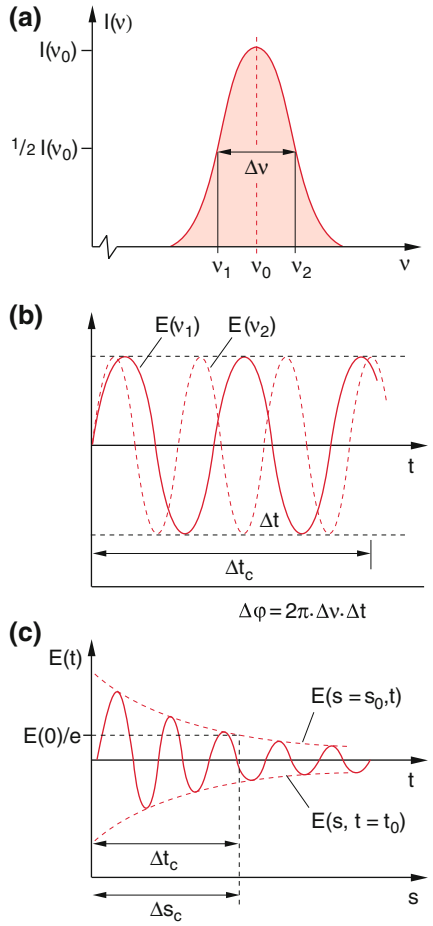


Fig. 10.1 Temporal coherence of light with the spectral width $\Delta\nu$. The coherence time is $\Delta t_c = 1/\Delta\nu$. **a**) Spectral distribution of the intensity $I(\nu)$, **b**) temporal superposition of two partial waves with slightly different frequencies, **c**) temporal decay of the total field amplitude $E(t)$ of all components in (**a**)

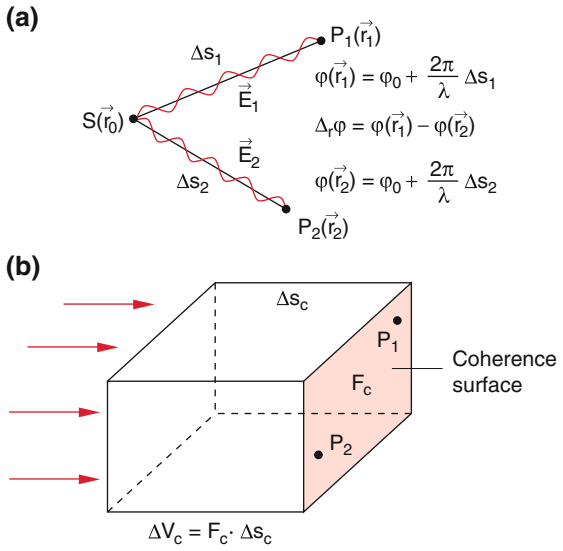


Fig. 10.2 **a**) Phase difference $\Delta_r\varphi$ between the phases $\varphi(\boldsymbol{r}_1)$ and $\varphi(\boldsymbol{r}_2)$ of a monochromatic wave at two different spatial points, **b**) coherence surface F_c and coherence volume ΔV_c

This can be also explained as follows:

The superposition of all spectral components results in a total amplitude $\boldsymbol{E}(t)$ which decays in time, because in the course of time more and more destructive interference with phase differences $\Delta\varphi > \pi/2$ appear. (Note that two waves with equal amplitude but a phase difference of π completely cancel each other).

The calculation shows that after the coherence time Δt_c the total amplitude has decreased to $1/e$ of its initial value (Fig. 10.1c).

Light with a spectral width $\Delta\nu$ therefore represents a finite wave train with the **coherence length** $\Delta s_c = c \cdot \Delta t_c$.

The coherence length Δs_c is the path length that light propagates during the coherence time Δt_c.

The phase differences $\Delta\varphi_{j,k}$ between the partial waves $\boldsymbol{E}_j$ and $\boldsymbol{E}_k$ may be different for different points $P(\boldsymbol{r})$ within the interference volume, because the phase differences $\Delta\varphi_i = (2\pi/\lambda_i)\Delta s_i$ of each partial wave E_i depends on the path difference $\Delta s_i = SP$ between source S and observation point P. If this spatial phase difference

$$\Delta_r\varphi_i = \varphi_i(\boldsymbol{r}_1) - \varphi_i(\boldsymbol{r}_2) \tag{10.4}$$

for an arbitrary partial wave $\boldsymbol{E}_i$ changes during the observation time Δt by less than 2π, the wave field is spatially coherent (Fig. 10.2). The surface perpendicular to the propagation direction is called the coherence surface A_c.

The product of coherence surface and coherence length Δs_c is the **coherence volume $\Delta\mathbf{V_c}$** [1].

Interference structure can be observed only within the coherence volume.

Example

1. For light with the spectral width $\Delta\nu = 2 \times 10^9$ Hz (typical Doppler width of spectral lines in the visible range) the coherence time is $\Delta t_c = 1/\Delta\nu = 5 \times 10^{-10}$s. The coherence length $\Delta s_c = c \cdot \Delta t_c$ is then $\Delta s_c = 0.15$ m.
2. A plane wave $\boldsymbol{E} = \boldsymbol{A} \cdot e^{i(\omega t - \boldsymbol{k}\cdot\boldsymbol{r})}$ is on the entire plane $\boldsymbol{k} \cdot \boldsymbol{r} = \text{const.}$ spatially coherent. If the wave is monochromatic.$(\Delta\nu = 0)$ the coherence length is infinite. Such a wave is coherent in the entire space. If its spectral width $\Delta\nu > 0$ is finite the coherence length is finite and the coherence volume is infinite in the direction perpendicular to the wave vector $\boldsymbol{k}$ but finite in the propagation direction.
3. A monochromatic spherical wave is coherent in the entire space. The general rule is: Waves emitted

from a point source (which is of course an idealization) are spatially coherent in the entire space.

We will now illustrate the terms *coherence* and *interference* by some examples which show how coherent waves can be generated and how their superposition can be realized.

10.2 Generation and Superposition of Coherent Waves

There are in principle two different methods how to generate coherent partial waves and their superposition, which results in interference structures:

- The emitters (i.e. the sources of the partial waves) are phase-locked with each other (Fig. 10.3).
- The waves emitted by the source S are split into two or more partial waves, which propagate through different paths with different path lengths until they are again superimposed. Their interference structure can then be observed in the points P_1 or P_2 (Fig. 10.4).

The first method can be realized for acoustic waves (see Vol. 1, Sect. 11.10) by feeding two spatially separated loudspeakers which are both driven by the same ac-voltage source.

In case of optical waves the sources are energetically excited atoms or molecules (see Vol. 3) which are generally independent of each other and emit light waves with randomly distributed phases. The light emitted from the whole light source is therefore incoherent. This implies that the first method is not applicable for classical light sources such as light bulbs, gas discharges or the sun.

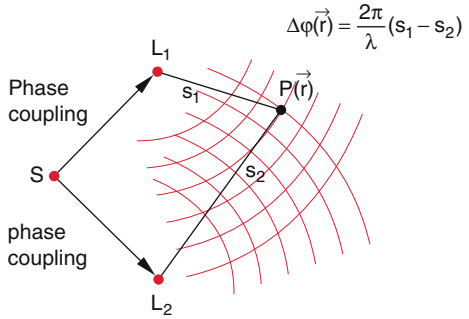


Fig. 10.3 Two sources L_1 and L_2, phase-locked to the common source S. The waves emitted by L_1 and L_2 superimpose within the coherence volume with temporally constant but location-dependent phase differences $\Delta\varphi(\boldsymbol{r})$ …

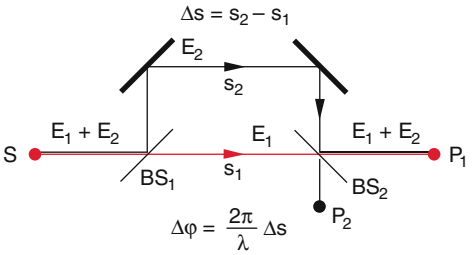


Fig. 10.4 Two-beam interference realized by splitting the incident wave into two partial waves which are again superimposed after having travelled different path lengths

Nowadays one can use frequency-stabilized lasers (see Vol. 3). Here the atoms can be excited by the coherent laser wave to phase-coupled oscillations (see Sect. 8.2). For instance in Fig. 8.4 all atoms in the plane $z = z_0$ oscillate in phase, if the incident wave is coherent.

With special techniques it is possible to phase-lock two stabilized lasers with each other. This enables one to use the first method for generating interference patterns also in the optical range.

Stabilized lasers are often used to demonstrate interference and diffraction phenomena to a large auditorium, because their coherence length is much larger than that of incoherent light sources.

Lasers are therefore often named "*coherent light sources*".

In most cases in optics the second method is preferred, even when lasers are used. The waves emitted by the source are split by different realizations of beam splitters, propagate through different path lengths and are again superimposed in the interference space.

Two beam interference occurs when two partial waves are superimposed, whereas **multiple beam interference** means the superposition of several partial waves.

Interference is the basis of all interferometers. These are devices for the accurate measurement of wavelengths. They are also applied to the very precise determination of small distance-changes or the recording of refractive indices of transparent media and their dependence on pressure and temperature.

Note Interference as spatially structured and temporal constant intensity $I(r)$ can be only observed in a limited volume of the superimposed partial waves where the path difference Δs between the partial waves is smaller than the coherence length $\Delta s_c = c \cdot \Delta t_c$. We will see, that the coherence volume depends on the spatial as well on the temporal coherence of the interfering waves.

10.3 Experimental Realization of Two-Beam Interference

There are a large number of possible experimental arrangements, which split the light wave from a source S into two partial waves, which can be again recombined by mirrors or lenses. We will illustrate this by some examples.

10.3.1 Fresnel's Mirror Arrangement

The light from a point source S is reflected by two mirrors M_1 and M_2 which are slightly tilted against each other (Fig. 10.5). For an observer in the observation plane (which we choose as the x-y-plane), the light reflected by the two mirrors seems to come from the two virtual sources S_1 and S_2.

The optical path lengths $s_1 = SM_1P(x, y)$ and $s_2 = SM_2P(x, y)$ is equal to the path lengths S_1P and S_2P. If the two virtual light sources have the coordinates ($x = \pm d$; $y = 0$, $z = z_0$) the path difference between S_1P and S_2P is

$$\Delta s = \sqrt{(x+d)^2 + y^2 + z_0^2} - \sqrt{(x-d)^2 + y^2 + z_0^2}. \quad (10.5)$$

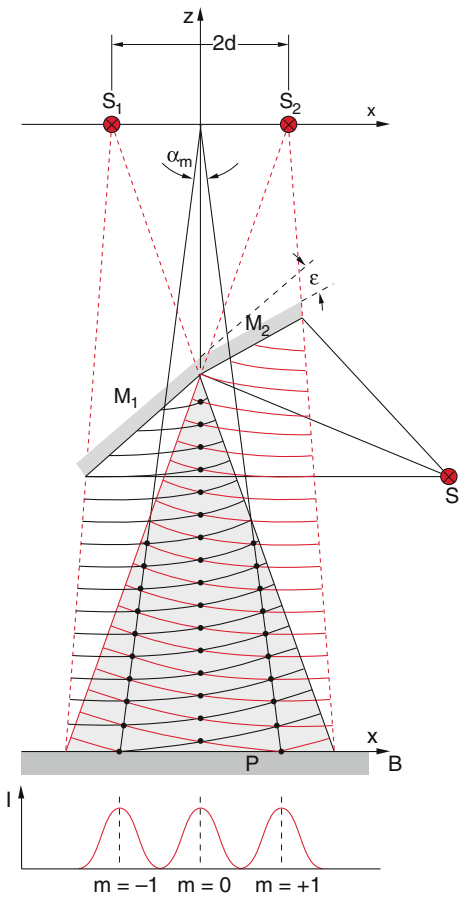


Fig. 10.5 Fresnel's mirror experiment

All points $P(x, y)$ for which Δs = const. form a hyperbola (see Problem 10.1).

For $\Delta s = m \cdot \lambda$ the two partial waves are in phase, i.e. they superimpose constructively and the maximum intensity

$$I_{\max} = c\varepsilon_0(\boldsymbol{E}_1 + \boldsymbol{E}_2)^2$$

is observed (black points in Fig. 10.5).

For $\Delta s = (2m+1)\lambda/2$ the two waves have opposite phases and the total intensity is

$$I_{\min} = c\varepsilon_0(\boldsymbol{E}_1 - \boldsymbol{E}_2)^2$$

One observes therefore in the x-y-plane a pattern of dark and bright hyperbolas. The spatial extension of the pattern is determined by the coherence length $\Delta s_c = c/\Delta\nu$ and therefore by the spectral width $\Delta\nu$ of the radiation from the light source and by the mutual distance of the two virtual light sources S_1 and S_2.

We have assumed a point like light source where the spatial extension of the source has been neglected. We will now discuss the influence of the finite size of the source on the extension of the coherence volume.

10.3.2 Young's Double Slit Experiment

The light emitted by an extended light source with diameter b illuminates two slits S_1 and S_2 which are separated by the distance d (Fig. 10.6). The total amplitude and the phase at the slits is obtained by the superposition of all partial waves emitted from the different surface elements dF_i of the source LS. For the determination of the phase one has to take into account the different path lengths from the elements dF_i to the slits S_1 and S_2.

The two slits can be regarded as sources of new waves which superimpose (Huygens's principle, see Vol. 1, Sect. 11.11). The total intensity at the point P in the observation plane is determined by the amplitudes A_i and the phases φ_i of the partial waves at the slits S_1 and S_2 and by the path difference $\Delta s = S_2P - S_1P$.

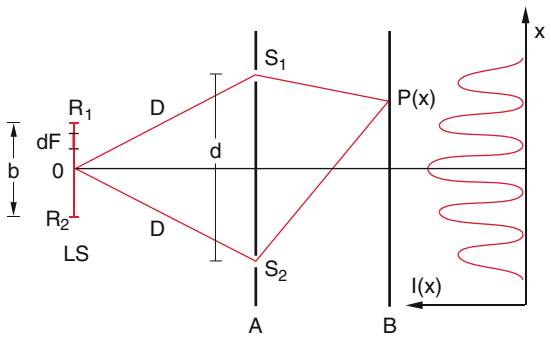


Fig. 10.6 Young's double-slit demonstration

If the different surface elements dF_i emit independently waves with randomly distributed phases (this is the case for incoherent sources), the phase of the total wave $E = \Sigma E_i$ will show a randomly fluctuating phase at S_1 and S_2. However, this would not affect the intensity at the point P as long as the fluctuations occur synchronously in S_1 and S_2 because then the phase difference $\Delta\varphi = \varphi_1 - \varphi_2$ of the waves emitted by the slits would be constant in time. In this case the two slits represent two coherent light sources which produce in the observation plane a static interference pattern, completely analog to that in Fresnel's mirror experiment.

For light emitted from the center O of the extended source this situation occurs indeed because the path lengths OS_1 and OS_2 are equal. Therefore phase fluctuations of the source arrive simultaneously at the two slits S_1 and S_2. This is no longer true for all other points Q of the light source, where path length differences $\Delta s = QS_1 - QS_2 \neq 0$ appear which are maximum for the points at the edge of the source (Fig. 10.7).

With the distance $D \gg d$ between source and the slits the maximum path difference is

$$\begin{aligned}\Delta s_{\max} &= R_1S_2 - R_1S_1 = R_2S_1 - R_1S_1 \approx b \cdot \sin\vartheta \\ &= 1/2 \cdot b \cdot d/(2D)\end{aligned} \tag{10.6}$$

Because of symmetry reasons is $R_1S_2 = R_2S_1$.

If for random emission of the different source elements Q the maximum path difference is larger than the half wavelength ($\Delta s_{\max} > \lambda/2$), the phase difference $\Delta\varphi = \varphi(S_1) - \varphi(S_2) = (2\pi/\lambda) \cdot \Delta s$ fluctuates by more than π and the interference structure in the observation plane disappears.

The condition for the coherent (i.e. phase correlated) illumination of the two slits by an extended source with diameter b is then

$$\begin{aligned}\Delta s_{\max} &\approx \frac{b \cdot d}{2D} < \lambda/2 \\ \Rightarrow \frac{d}{\lambda} &< \frac{D}{b} \Rightarrow \frac{d^2}{\lambda^2} < \frac{D^2}{b^2} \approx \frac{1}{\Delta\Omega}.\end{aligned} \tag{10.7}$$

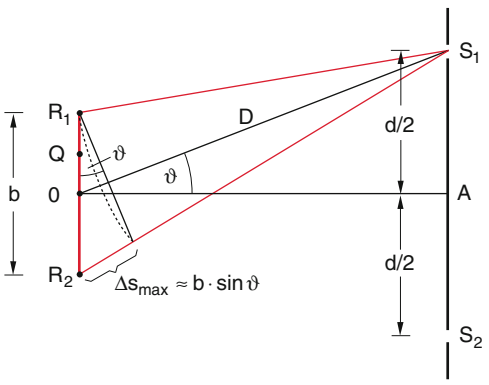


Fig. 10.7 The influence of the source size on the coherence properties at the two slits

where $\Delta\Omega$ is the solid angle under which the light source LS is seen from the point A between the two slits.

For coherent illumination of two slits by an extended light source the distance d/λ of the slits in units of the wavelength λ should not exceed the ratio D/b of distance D from the slits to the source and the diameter b of the source.

Since $b^2/D^2 = \Delta\Omega$ is the solid angle under which the source area b^2 appears from the slits, we can also formulate this condition as follows:

The coherence surface d^2/λ^2 in units of λ^2 is equal to the solid angle $\Delta\Omega$.

The coherence surface of the radiation from an extended source is

$$A_c = d^2 \leqq \lambda^2/\Delta\Omega$$

where $\Delta\Omega$ is the solid angle under which the source appears from a point of the coherence surface.

If the condition (10.7) is satisfied, an interference pattern can be observed in the observation plane, even for the illumination by an extended incoherent source. The extension of the source can be the larger the farther away it is.

Examples

1. $b = 1$ cm, $D = 50$ cm, $\lambda = 500$ nm, $\Rightarrow d \leq 25$ µm.
2. Our next fixed star (besides our sun) is Proxima Centauri. Its distance is $D = 4.3\ Ly = 4 \times 10^{16}$ m and its diameter $b \approx 10^{10}$ m. The diameter d of the coherence surface on earth is then for $\lambda = 500$ nm $d \approx 2$ m.

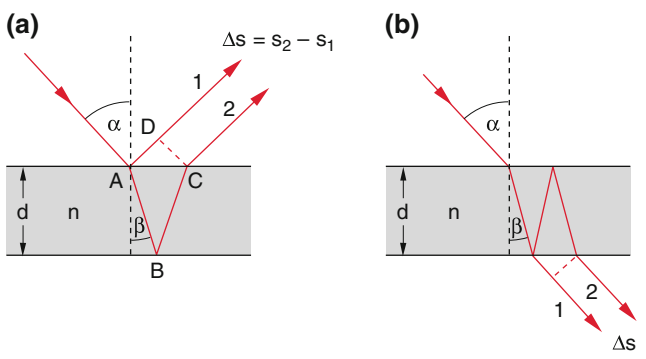


Fig. 10.8 Determination of the path difference between the two partial waves reflected at the two surfaces of a plane parallel transparent plate. **a**) For the reflected light, **b**) for the transmitted light

10.3.3 Interference at a Plane-Parallel Plate

When a plane wave falls under the angle α onto a transparent plane-parallel plate with refractive index n (Fig. 10.8) part of the wave is reflected and the other part is refracted (see Sect. 8.4). The refracted wave is again reflected at the lower surface, is refracted at the upper surface and leaves the plate parallel to the partial wave 1.

The path difference between the two partial waves is for a thickness d of the plate according to Fig. 10.8

$$\begin{aligned}\Delta s &= n \cdot (\overline{AB} + \overline{BC}) - \overline{AD} \\ &= \frac{2nd}{\cos\beta} - 2d \tan\beta \sin\alpha.\end{aligned}$$

With $\sin\alpha = n \cdot \sin\beta$ this can be written as

$$\begin{aligned}\Delta s &= \frac{2nd}{\cos\beta} - \frac{2nd \sin^2\beta}{\cos\beta} = 2nd \cos\beta \\ &= 2d\sqrt{n^2 - \sin^2\alpha}.\end{aligned} \tag{10.8}$$

Since a phase jump of π occurs for the reflection at the upper surface, (see Sect. 8.4.8) the phase difference between the two reflected partial waves is

$$\Delta\varphi = \frac{2\pi}{\lambda}\Delta s - \pi. \tag{10.9}$$

The two partial waves interfere constructively (their amplitudes add) for $\Delta\varphi = m \cdot 2\pi$ whereas for $\Delta\varphi = (2m+1)\pi$ minimum intensity is observed (destructive interference, the amplitudes are subtracted).

When the plate is illuminated by divergent monochromatic light with wavelength λ which contains rays with angles of incidence α in the range $\alpha_0 \pm \Delta\alpha$ one obtains for all angles α maximum intensity if

$$2d\sqrt{n^2 - \sin^2\alpha} = (m + 1/2)\lambda. \tag{10.10}$$

One therefore observes in the reflected light a system of bright and dark concentric rings around the normal to the surface (Fig. 10.9).

Also for the transmitted light the path difference between two partial waves is given by (10.8) as can be readily verified.

However, here no phase jump occurs. The phase difference is therefore instead of (10.9) now $\Delta\varphi = (2\pi/\lambda) \cdot \Delta s$. Maximum transmission is obtained for $\Delta s = m \cdot \lambda$. The reflected intensity is then minimum.

For small reflectivity $R \ll 1$ (for instance for an uncoated glass plate) the influence of the multiple reflected partial waves can be neglected and we have an example of two-beam interference. It can be demonstrated for a large auditorium as shown in Fig. 10.9 where a thin glass plate was illuminated with the divergent beam from an argon ion laser. Because of the high intensity of the laser the entire wall of the lecture hall can be filled with such an interference pattern.

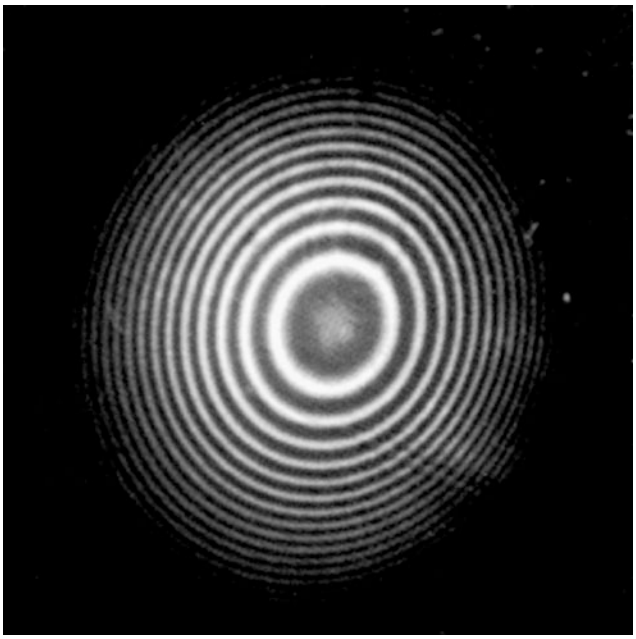


Fig. 10.9 Photo of the interference rings observed with the divergent green light of an argon laser reflected by a plane parallel glass plate

10.3.4 Michelson Interferometer

We regard in Fig. 10.10 a parallel light beam which propagates into the z-direction. It is split by the beam splitter BS into two partial beams. The reflected part travels into the y-direction, is reflected at the mirror M_1 and transmitted through BS and reaches the observation plane B. The second partial beam is transmitted by BS, reflected by the movable mirror M_2, is again reflected at BS and superimposes the first partial beam in the observation plane B. For strictly plane waves and correctly aligned mirrors the observation plane is a phase plane. Therefore for destructive interference it is completely dark, for constructive interference it appears bright without any spatially structured interference pattern.

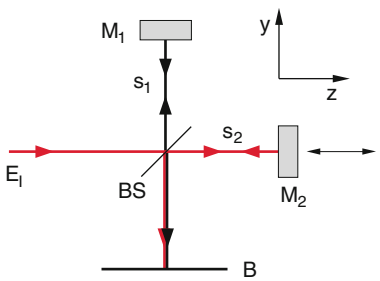


Fig. 10.10 Schematic drawing of the Michelson interferometer

We will now determine the intensity I_t transmitted by the interferometer as a function of the path difference $\Delta s = s_1 - s_2$.

The incident plane wave

$$\boldsymbol{E}_\mathrm{i} = \boldsymbol{A}_\mathrm{i} \cdot \cos(\omega t - kz). \tag{10.11a}$$

is reflected at M_1 (reflectivity R) and transmitted through the beam splitter BS (transmission T), In the observation plane it has therefore the amplitude

$$|\boldsymbol{E}_1| = \sqrt{R \cdot T} A_\mathrm{i} \cdot \cos(\omega t + \varphi_1), \tag{10.11b}$$

where the phase φ_1 depends on the optical path $BS - M_1 - BS - B$.

For the second partial wave we obtain

$$|\boldsymbol{E}_2| = \sqrt{R \cdot T} A_\mathrm{i} \cdot \cos(\omega t + \varphi_2). \tag{10.11c}$$

where the phase φ_2 depends on the path BS-M_2-BS-B. The amplitude of the two waves in the plane B is equal, independent of the reflectivity R or the transmission T of the beam splitter BS, because both waves suffer one reflection and one transmission.

Note This is not true for the wave reflected back into the source.

The total transmitted intensity observed in the plane B is then

$$\begin{aligned} I_T &= c\varepsilon_0 (E_1 + E_2)^2 \\ &= c\varepsilon_0 RTA_\mathrm{i}^2 [\cos(\omega t + \varphi_1) + \cos(\omega t + \varphi_2)]^2. \end{aligned} \tag{10.12}$$

Since the detector in B cannot follow the rapid optical oscillations with frequency ω, it averages over many oscillation period $T = 2\pi/\omega$. We therefore obtain from (10.12) with $\langle \cos^2 \omega t \rangle = 1/2$ the time average of the transmitted intensity

$$I_\mathrm{T} = RTI_0(1 + \cos \Delta\varphi), \tag{10.13}$$

where $I_0 = c \cdot \varepsilon_0 \cdot E_\mathrm{i}^2$ is the incident intensity and $\Delta\varphi = \varphi_1 - \varphi_2 = (2\pi/\lambda) \cdot \Delta s$ is the phase difference between the two interfering waves, which depends on the path difference $\Delta s = s_1 - s_2$ and the wavelength $\lambda = 2\pi c/\omega$ of the incident wave. For $R = T = 0.5$ we obtain for the temporal average of the intensity with $I_T = \frac{1}{2} I_0$

$$I_\mathrm{T} = \frac{1}{2} I_0 (1 + \cos \Delta\varphi). \tag{10.13a}$$

Depending on the phase difference the transmitted intensity varies between zero (for $\cos \Delta\varphi = -1$) and I_0 for $\cos \Delta\varphi = +1$ (Fig. 10.11). For $I_T = 0$ $(\Delta = (2m+1) \cdot \pi)$ the total intensity is reflected back into the source.

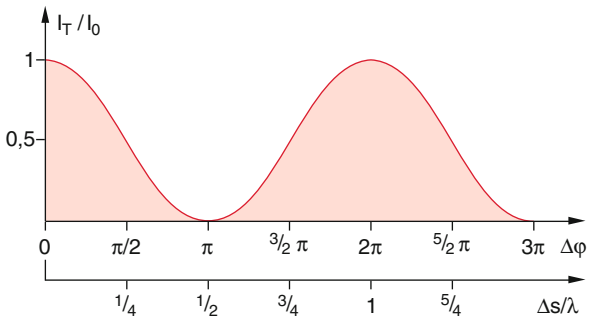


Fig. 10.11 Transmitted intensity of the Michelson interferometer as a function of the path difference between the two arms

The Michelson interferometer with $R = T = 0.5$ acts as a wavelength dependent mirror. If the incident light has a broad spectral bandwidth, all wavelengths $\lambda_\mathrm{m} = 2\Delta s/(2m+1)$ are completely reflected, whereas the wavelengths $\lambda_\mathrm{m} = \Delta s/m$ are completely transmitted. All wavelengths in between these limiting cases are partly reflected and partly transmitted.

The Michelson interferometer can be used for very precise wavelength measurements. The mirror M_2 is placed on a carriage which smoothly moves with the velocity v on an air track. The path difference between the two partial waves

$$\Delta s(t) = \Delta s(0) + 2v \cdot t \tag{10.14}$$

is now time dependent and a constant sequence of intensity maxima and minima are observed in the plane B. After a time $T = N \cdot \lambda/2v$ where the mirror has moved over the path length Δz the number of measured intensity maxima is N. The wavelength is then

$$\lambda = 2v \cdot T/N = 2\Delta z/N$$

Example

With a velocity $v = 0.1$ m/s and an observation time $T = 10$ s the carriage travels the distance $\Delta z = 1$ m. The number of counts N is for a wavelength $\lambda = 500$ nm $N = 4 \times 10^6$. The relative uncertainty

$$\Delta\lambda/\lambda = 2\Delta z/\lambda(1/N - 1/(N+1)) = 2.5 \times 10^{-7}.$$

Of course the path difference should not be larger than the coherence length. The example above is for wavelength measurements of stabilized lasers with a coherence length of more than 1 m.

The intensity contrast between the maximum and minimum intensity of the transmitted light

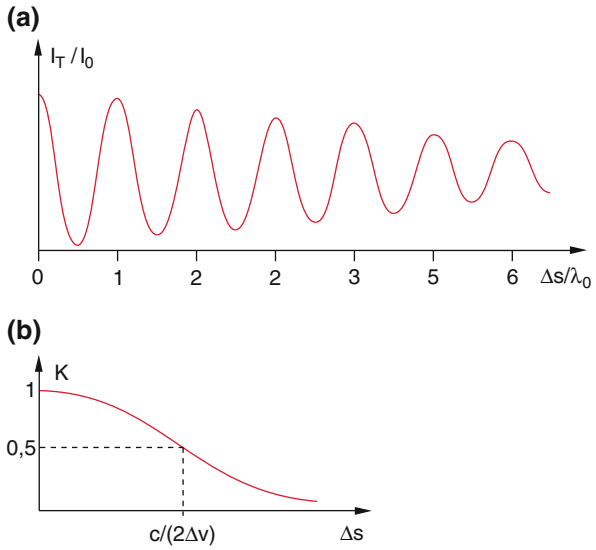


Fig. 10.12 **a**) Transmitted intensity of the Michelson interferometer of a Doppler-broadened spectral lines as a function of the path difference Δv. **b** Contrast function $K(\Delta s, \Delta v)$ for light with a spectral Gauß-profile with spectral width Δv

$$K = (I_{\max} - I_{\min})/(I_{\max} + I_{\min})$$

approaches zero for $\Delta z \gg \Delta s_c$ with increasing Δz (Fig. 10.12).

Example

For a spectral bandwidth $\Delta\nu \leq 3 \times 10^9\ s^{-1}$ of the incident radiation the coherence length becomes $\Delta s_c = c/\Delta\nu \geq 0.1$ m. A path difference of $2\Delta s = 0.2$ m can be realized when the carriage moves over a distance of 10 cm in such a way that Δs changes from −5 to +5 cm. The absolute uncertainty for the wavelength measurement is than $\Delta\lambda = 1.25 \times 10^{-3}$ nm. Since the light from stabilized lasers has a coherence length of $\Delta s_c > 5$ m, one can realize path differences of $\Delta s > 5$ m. With modern devices a relative accuracy of better than $\Delta\nu/\nu = 10^{-8}$ can be achieved [2, 3].

If the incident light beam is strictly parallel but the mirrors of the Michelson interferometer are slightly tilted, one observes in the plane B a regular system of bright and dark parallel lines.

In practice strictly parallel light beams are difficult to realize and the incident light is generally slightly divergent. The light rays in the divergent beam have slightly different inclination angles against the z-axis (Fig. 10.13). Since the path difference $\Delta s = \Delta s_0/\cos\alpha$ depends on the inclination angel α one obtains in the plane B not a uniform intensity, independent of x and y as for parallel incident light, but a

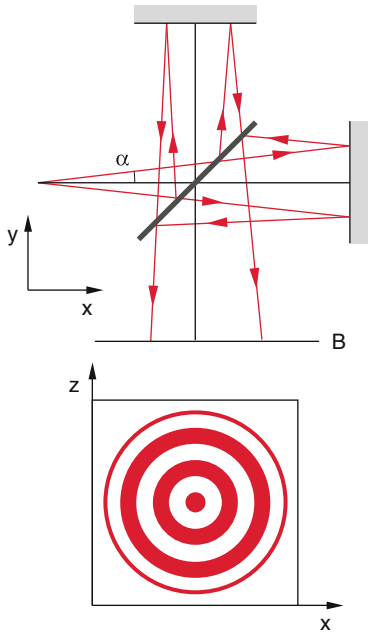


Fig. 10.13 Generation of a ring system for divergent incident light

system of concentric bright and dark rings (see Sect. 10.3.3 and Fig. 10.9). For the bright rings the condition $\Delta s = m\lambda$ is fulfilled, for the dark rings is $\Delta s = (2m + 1) \cdot \lambda/2$.

10.3.5 The Michelson-Morley Experiment

Such an interferometer was used 1887 by A. Michelson (Fig. 10.14) and E. Morley in order to clarify whether there is a medium, that fills the entire space of the universe, where the electromagnetic waves penetrate through this medium and which possibly rests relative to the earth. This medium,

Fig. 10.14 *Albert Abraham Michelson* (1852–1931) (with kind permission of Deutsches Museum München)

which was controversial discussed among scientists, was called "*ether*". The experiment should determine the relative motion of the earth against this ether and should resolve the problem whether the speed of light c depends on the direction of the incident light against or with the velocity $\boldsymbol{v}$ of the earth moving around the sun (see Vol. 1, Sect. 3.4), which would be true if the ether rests relative to the moving earth. This can be illustrated by the example of water waves:

When throwing a stone from a moving boat into the water spherical waves are generated. The phase velocity of these waves is independent of the direction in a coordinate system of an observer resting against the water. However, relative to the boat, which has the velocity v_S against the water, the phase velocity of the waves in the driving direction of the boat is $v_1 = v_{ph} - v_S$ and in the opposite direction $v_2 = v_{ph} + v_S$. Measuring the two velocities v_1 and v_2, the velocity $v_{ph} = (v_1 + v_2)/2$ of the boat as well as the phase velocity $v_S = (v_2 - v_1)/2$ of the water wave can be determined. The question is now whether the same situation applies for light waves.

By corresponding experiments with light from a star the two scientists hoped to be able to determine the speed of light c and the velocity v of the earth relative to the ether. Many earlier experiments by Fizeau, Michelson and other scientists had already proved that the moving earth did not take along a possible ether. This means when the earth moves with the velocity v, the ether has the velocity $-v$ against the earth if it rests relative to the sun.

When the Michelson interferometer is orientated in such a way that one arm is parallel, the other perpendicular to the velocity $\boldsymbol{v}$ of the earth (Fig. 10.15), the time of flight of the light from the beam splitter BS to the mirror M_1 and back to BS, should be for the parallel arm with length L_1

$$t_{||} = \frac{L}{c-v} + \frac{L}{c+v} = \frac{2cL}{c^2 - v^2} = \gamma^2 \frac{2L}{c} \quad \text{with} \quad \gamma = \left(1 - \frac{v^2}{c^2}\right)^{-1/2}. \tag{10.15a}$$

because the light travels on its way BS-M_1 against the ether and therefore should have the velocity $c - v$ and on its way back $c + v$.

For the arm perpendicular to $\boldsymbol{v}$ the mirror M_2 moves during the time of flight t_2 along the distance $\Delta z = v \cdot t_2$. The light beam therefore has to be inclined against the vertical arm in order to reach the mirror M_2. The inclination can be calculated by vector addition of the two perpendicular distances $\boldsymbol{L} + \boldsymbol{v} \cdot t_2 = \mathrm{c} \cdot t_2$ (Fig. 10.15c). We get the relation

$$c^2 t_2^2 = v^2 t_2^2 + L^2,$$

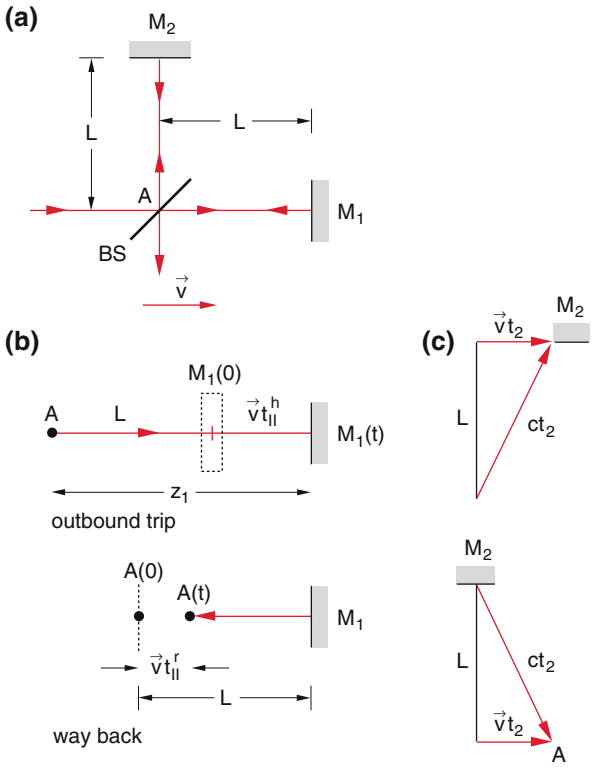


Fig. 10.15 Experimental examination of a possible time difference between the two partial waves in the Michelson interferometer if there were a resting ether. **a**) Schematic experimental arrangement, **b**) time diagram for the travel time of light parallel to the velocity v of the earth, **c**) perpendicular to v

This gives for the time of flight back and forth between BS and M_2

$$t_\perp = \frac{2}{\sqrt{c^2 - v^2}} = \gamma \cdot \frac{2L}{c} \tag{10.15b}$$

The time difference between the two partial waves is then

$$\Delta t = t_{||} - t_\perp = \frac{2L}{c}\left(\gamma^2 - \gamma\right). \tag{10.16}$$

The velocity v of the earth on its way around the sun is $v \approx 3 \times 10^4 \mathrm{m/s}$ (the additional velocity due to the rotation of the earth around its axis is at the latitude $\varphi = 45°$ only $3.2 \times 10^2 \mathrm{m/s}$, i.e. 1% of v, and can be therefore neglected).

We can therefore approximate

$$\gamma \approx 1 + \frac{1}{2} v^2/c^2 \quad \text{and} \quad \gamma^2 = 1 + v^2/c^2.$$

This converts (10.16) to

$$\Delta t = L \frac{v^2}{c^3}. \tag{10.17a}$$

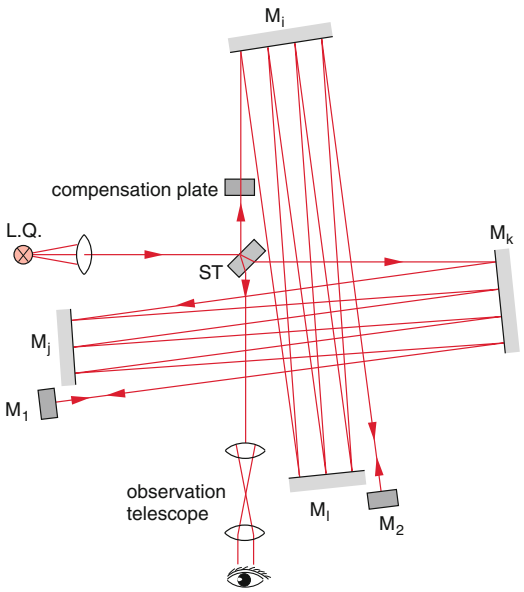


Fig. 10.16 Schematic arrangement of the Michelson-Morley experiment The mirrors M_1 and M_2 are adjustable to reflect the light beams back into themselves

Δt is the time difference that corresponds to a phase difference

$$\Delta\varphi = 2\pi v \Delta t = \frac{2\pi c}{\lambda}\Delta t \approx \frac{2\pi}{\lambda}\frac{Lv^2}{c^2}. \qquad (10.17b)$$

For incident light which is slightly inclined against the interferometer axis (this means that a light beam in y-direction is slightly inclined against the z-axis) interference stripes appear in the observation plane (see Problem 10.3) which can be observed with a magnifying telescope with cross hairs (Fig. 10.16). A phase difference $\Delta\varphi$ corresponds to the shift of x interference stripes in the observation plane B, where

$$x = \frac{\Delta\varphi}{2\pi} = \frac{Lv^2}{\lambda c^2}. \qquad (10.17c)$$

When the whole interferometer which is mounted on a turn table (Fig. 10.17) is rotated by 90° a stripe shift

$$\Delta m = 2x = \frac{2Lv^2}{\lambda c^2} \qquad (10.18)$$

should be expected if the ether theory is valid.

Michelson and Morley increased the sensitivity of their interferometer by multiple reflections in both arms (Fig. 10.16) which allowed them to realize an effective length L = 11 m. Inserting the numerical values $L = 11$ m; $v^2/c^2 = 10^{-8}$ and $\lambda = 5 \times 10^{-7}$ m yields a stripe shift of $\Delta m = 0.4$. This is far above the detection limit.

For the real experiment the researchers used the following tricks:

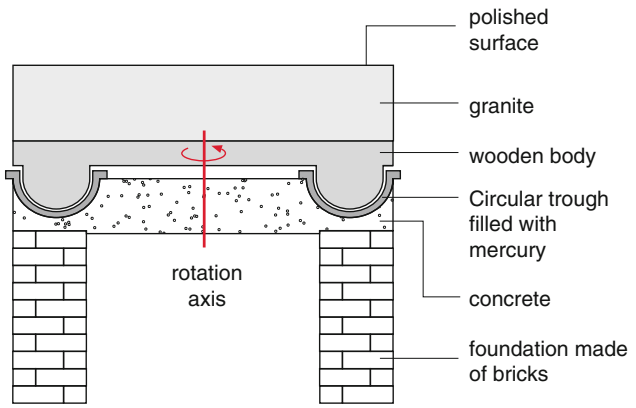


Fig. 10.17 Experimental realization of the rotatable interferometer

- Since the star light has a large spectral width $\Delta\lambda$ the transmission through the beam splitter BS causes dispersion which results in a washing out of the interference stripes because the phase shifts depend on the wavelength λ. This can be avoided by inserting a compensation plate in one arm of the interferometer which compensates the dispersion in the other arm, because now the light in both arms has to pass through a glass plate.
- The whole basis of the interferometer rests on a stone plate which in turn is placed on a wooden body that floats on a mercury bath (Fig. 10.17). This allows the easy rotation of the whole system [5].

In spite of very careful measurements which were several times repeated no stripe shift could be detected when the interferometer was turned by 90°. Therefore Michelson and Morley concluded correctly, that the speed of light is equal for all directions and independent of the velocity of the detector or the light source (see Vol. 1, Sect. 3.4). **This also means that there is no ether** [4]. Its existence is furthermore not at all necessary for the explanation of the observations, as is evident from the theory of electromagnetic waves discussed in more detail in Chap. 6, which shows that electromagnetic waves can propagate also in vacuum and do not need any material such as the ether for their propagation.

10.3.6 Sagnac Interferometer

Similar to the Michelson interferometer also the *Cagnac interferometer* has proved to be an important tool for tests of General Relativity and also for navigation purposes. Its principle is illustrated in Fig. 10.18. The incident plane wave with intensity $I = I_0 \cdot \cos\,\omega t$ is split by the beam splitter *BS* into two partial waves which circumvent the area A in

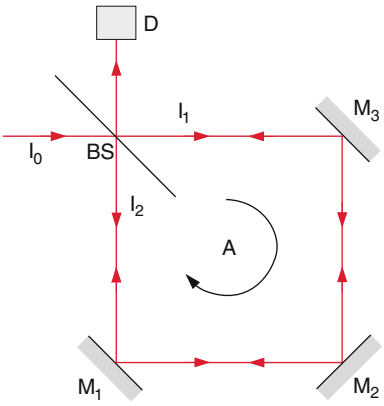


Fig. 10.18 Sagnac interferometer

opposite directions: One wave with intensity I_1 traverses the path BS-M_3-M_2-M_1-BS clockwise, the other wave with intensity I_2 anticlockwise. Both waves are superimposed at BS and reach the detector D. If the interferometer is at rest (no rotation) both waves are in phase and interfere constructively. The total intensity is then $I = I_1 + I_2 = I_0$.

If the whole interferometer rotates, for instance clockwise, then the wave in clockwise direction has to pass a longer way (the mirrors recede for this wave) than the anticlockwise wave (the mirrors move against the wave). Therefore a phase difference $\Delta\varphi$ appears between the two waves at BS and the total intensity at the detector D is now for $I_1 = I_2 = I_0/2$

$$I(\Delta\varphi) = I_1 + I_2 \cos\Delta\varphi = \frac{1}{2} I_0 (1 + \cos\Delta\varphi),$$

where the phase difference

$$\Delta\varphi = \frac{8\pi A}{c \cdot \lambda} \Omega \cos\Theta$$

depends on the area A, the angular rotation frequency Ω of the interferometer, the wavelength λ of the incident light and the inclination angle Θ between rotation axis and the normal of the area A.

From the measured fringe shift $\Delta = \Delta\varphi/2\pi$ the phase difference $\Delta\varphi$ and the rotation frequency Ω is obtained.

Using such a Cagnac interferometer with an enclosed area $A = 207{,}836\,\mathrm{m}^2$ Michelson and Gale have measured 1925 the earth rotation. They obtained an interference stripe shift of $\Delta = 0.230 \Rightarrow \Delta\varphi = \Delta \cdot 2\pi$ rad = 1.45 rad, which gives a rotation frequency $\Omega = 7.42 \times 10^{-5}\ s^{-1}$ of the earth. The correct value, obtained from the known rotation period of the earth is Δ = 1.48 rad, which gives a relative error of 0.2%. The inclination angle $\Theta = \vartheta$ for an interferometer with a horizontal area A depends on the latitude ϑ of the lab location.

Nowadays Sagnac interferometers are used with lasers as light sources (see Vol. 3) and optical fibers which circumvent many times the area A.

Example

With an area $A = 1\ \mathrm{m}^2$ and an optical fiber which circumvents the area A $N = 10^5$ times the effective area is $A_{\mathrm{eff}} = 10^5$ m. With stabilized lasers a phase difference $\Delta\varphi = 10^{-4}$ rad can be still measured. This allows a lower detection limit for the rotation frequency with an uncertainty $\Delta\Omega/\Omega = 5 \times 10^{-4}$.

For navigation purposes such a laser gyro consists of three Sagnac interferometers with areas that are mounted perpendicular to each other. Since the phase shift $\Delta\varphi$ depends on the latitude ϑ the position of a ship can be determined if the longitude is measured by accurate time measurements. However, nowadays the GPS navigation system (see Vol. 1) has replaced most Sagnac systems because it is more accurate and faster.

10.3.7 Mach-Zehnder Interferometer

In A Mach-Zehnder-interferometer the incident wave is split by the beam splitter BS_1 into two partial waves (Fig. 10.19). One passes through a cell that contains a medium with refractive index n and length L the other through open air. At the beam splitter BS_2 the two waves are again superimposed and reach either the detector D_1 or the detector D_2. The phase difference between the two waves depends on the path difference

$$\Delta s = \Delta s_0 + (n - 1) \cdot L$$

where Δs_0 is the path difference for the empty cell (n = 1).

When the gas pressure p in the cell is continuously changed, the path difference changes accordingly and the measured interference intensity $I(p)$ at the detectors D_1 and D_2 allows the accurate determination of the refractive index $n(p)$. One just counts the interference maxima $N = [n(p_1) - n(p_2)] \cdot L/\lambda$ which appear when the pressure is altered from p_1 to p_2 where the phase difference changes by

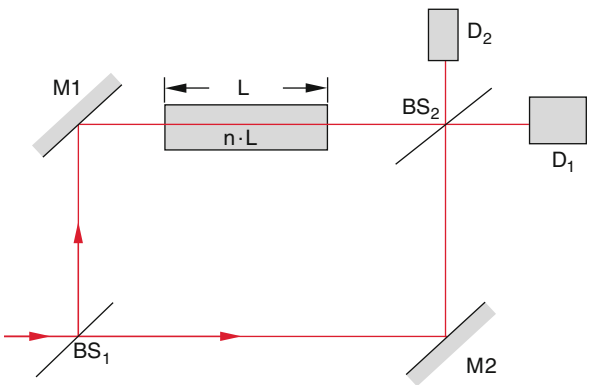


Fig. 10.19 Mach-Zehnder interferometer

$$\Delta\varphi = (2\pi/\lambda) \cdot L \cdot [n(p_1) - n(p_2)]$$

The phase difference at the detector D_2 differs by $\Delta\varphi_0 = \pi$ than that at D_1 because of the phase jump for the reflection at the optically thicker medium BS_2.

10.4 Multiple Beam Interference

Often interference phenomena are due to the superposition of many partial waves. For instance, if the two surfaces of the plane-parallel plate discussed in Sect. 10.3.3 are coated with highly reflecting layers the incident wave can be reflected many times between front surface and back surface (Fig. 10.20). All transmitted as well as all reflected partial waves can interfere.

We will now discuss this case in more detail, because it plays an important role in modern optics and interferometry. The discussion is similar to that in Sect. 10.3.3 but here the decrease of the amplitudes of the partial waves has to be taken into account.

When the incident plane wave

$$\boldsymbol{E} = \boldsymbol{A}_0 \cdot e^{i(\omega t - \boldsymbol{kr})}$$

Falls under the angle α onto the plane parallel plate, the amplitude A_i is split at both surfaces into a transmitted and a reflected part.

The reflected part has the amplitude $A_i \cdot \sqrt{R}$, while the transmitted part has the amplitude $A_i \cdot \sqrt{1-R}$), as long as absorption can be neglected. From Fig. 10.20 we obtain the following relations for the amounts of the amplitudes A_i of the waves reflected at the upper surface, B_i of the waves refracted at the upper surface, C_i of the waves reflected at the lower surface and D_i of the waves transmitted at the lower surface.

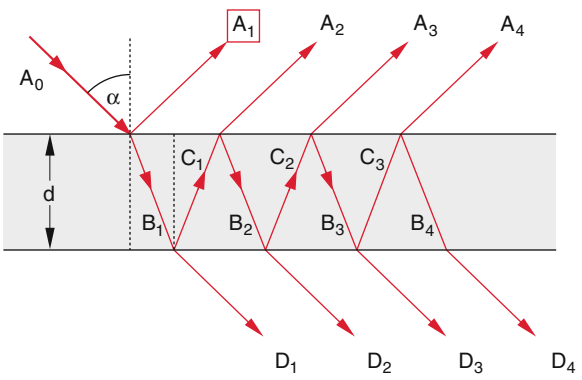


Fig. 10.20 Multiple beam interference at two plane-parallel flat surfaces with reflectivity R and distance d

$$\begin{aligned}
&|A_1| = \sqrt{R}|A_0|, \quad |B_1| = \sqrt{1-R}|A_0|,\\
&|C_1| = \sqrt{R(1-R)}|A_0|, \quad |D_1| = (1-R)|A_0|;\\
&|A_2| = \sqrt{1-R}|C_1| = (1-R)\sqrt{R}|A_0|,\\
&|B_2| = \sqrt{R}|C_1| = R \cdot \sqrt{1-R}|A_0|,\\
&|A_3| = \sqrt{1-R}|C_2| = R^{3/2}(1-R)|A_0| \quad \text{etc.}
\end{aligned} \tag{10.19}$$

The general expression for the amplitudes A_i with $i > 1$ is

$$A_{i+1} = R \cdot |A_i| \tag{10.20a}$$

And for the transmitted waves

$$D_{i+1} = R \cdot |D_i|. \tag{10.20b}$$

As has been shown in Sect. 10.3.3 the optical path difference between consecutive reflected as well as transmitted partial waves is

$$\Delta s = 2d\sqrt{n^2 - \sin^2\alpha},$$

This results in a phase difference

$$\Delta\varphi = 2\pi\Delta s/\lambda + \delta\varphi$$

where $\delta\varphi$ describes possible phase jumps on reflections at the medium with higher refractive index.

In Sect. 8.4.8 it has been shown that $\delta\varphi$ depends on the direction of the electric vector $\boldsymbol{E}$ of the incident wave which can be parallel or perpendicular to the plane of incidence.

For $E_\perp$ the following statements apply

(a) Under reflection at the optically thicker medium is $\delta\varphi = \pi$
(b) Under reflection at the optically thinner medium is $\delta\varphi = 0$.

For $\boldsymbol{E}_\parallel$ applies:

(a) Under reflection at the optically thicker medium $\delta\varphi = 0$ for all angles of incidence $\alpha < \alpha_B$, but $\delta\varphi = \pi$ for $\alpha > \alpha_B$, where α_B is the Brewster angle with ($\tan \alpha_B = n_2/n_1$).
(b) Under reflection at the optically thinner medium is $\delta\varphi = \pi$ for $\alpha < \alpha_B$ but $\delta\varphi = 0$ for $\alpha_B < \alpha < \alpha_c$ where α_c is the critical angle of total reflection ($\sin \alpha_c = n_2/n_1$).

For vertical incidence ($\alpha = 0$) the distinction between $\boldsymbol{E}_\perp$ and $\boldsymbol{E}_\parallel$ disappears. There is always a phase jump $\delta\varphi = \pi$ under reflection at the optically thicker medium and $\delta\varphi = 0$ at the optically thinner medium.

As can be derived from Fig. 10.20 the phase difference between the reflected waves A_i and A_{i+1} (for $i > 1$) is for all cases

$$\Delta\varphi = 2\pi\Delta s/\lambda.$$

Possible phase jumps under reflection influence the phase difference only for the transitions $A_0 \to A_1$ and $A_1 \to A_2$.

For $A_{1\perp}$ we get

$$A_{1\perp} = \sqrt{R} \cdot A_0 \cdot e^{i\pi} = -\sqrt{R}A_0, \tag{10.21a}$$

For $A_{1\parallel}$ is

$$A_{1||} = \pm\sqrt{R} \cdot A_0 \tag{10.21b}$$

where the sign depends on the condition $\alpha < \alpha_B$ (+sign) or $\alpha > \alpha_B$ (−sign).

The total amplitude of the reflected wave is the sum of all reflected partial waves where the phases have to be taken into account.

$$\begin{aligned} A &= \sum_{m=1}^{p} A_m e^{i(m-1)\Delta\varphi} \\ &= \pm A_0\sqrt{R} \cdot \left[1 - (1-R)e^{i\Delta\varphi} - R(1-R)e^{-2i\Delta\varphi} - \ldots\right] \\ &= \pm A_0\sqrt{R} \cdot \left[1 - (1-R)e^{i\Delta\varphi} \cdot \sum_{m=0}^{p-2} R^m e^{im\Delta\varphi}\right]. \end{aligned} \tag{10.22}$$

If the cross section of the plane parallel plate is sufficiently large or the incidence angle α sufficiently small, many partial waves can overlap and we can set $p \to \infty$. The limit $p \to \infty$ of the geometrical series (10.22) is

$$A = \pm A_0\sqrt{R}\frac{1 - e^{i\Delta\varphi}}{1 - Re^{i\Delta\varphi}}. \tag{10.23}$$

The intensity of the reflected wave is then

$$I_R = c\varepsilon_0 AA^* = I_0 \cdot R \cdot \frac{2 - 2\cos\Delta\varphi}{1 + R^2 - 2R\cos\Delta\varphi}.$$

This can be written with $1 - \cos x = 2\sin^2(x/2)$ as

$$I_R = I_0 \cdot \frac{4R \cdot \sin^2(\Delta\varphi/2)}{(1-R)^2 + 4R \cdot \sin^2(\Delta\varphi/2)}. \tag{10.24}$$

In a similar way one finds for the intensity of the transmitted wave

$$I_T = I_0 \cdot \frac{(1-R)^2}{(1-R)^2 + 4R \cdot \sin^2(\Delta\varphi/2)}. \tag{10.25}$$

From (10.24) and (10.25) we see that $I_R + I_T = I_0$, because we have neglected any absorption.

With the abbreviation

$$F = \frac{4R}{(1-R)^2}$$

We obtain from (10.24) and (10.25) the **Airy Formulas** for the reflected and transmitted intensities

$$I_R = I_0 \frac{F \cdot \sin^2(\Delta\varphi/2)}{1 + F\ \sin^2(\Delta\varphi/2)}. \tag{10.24a}$$

$$I_T = I_0 \frac{1}{1 + F\ \sin^2(\Delta\varphi/2)}, \tag{10.25a}$$

Since both intensities depend on the phase difference $\Delta\varphi$ between consecutive partial waves it is of interest to find experimental ways to alter $\Delta\varphi$. There are two ways to realize this:

(a) By tuning the wavelength λ for a fixed path difference $\Delta s = (\lambda/2\pi)\Delta\varphi$

(b) By variation of Δs for a fixed wavelength λ. This can be realized with the interferometer of Fig. 10.21b which consists of two plates each with one reflecting and one anti-reflecting surface layer. The two reflecting surfaces oppose each other and are carefully aligned to form a parallel air space between them.

For the case (a) a solid plane parallel plate with reflecting surfaces can be used (Fig. 10.21a). The incident light is either monochromatic and the wavelength λ is tuned, or a light source emitting a spectral continuum that contains all wavelengths between λ_1 and λ_2 is used. The interferometer then filters those wavelengths $\lambda_m = \Delta s/m$ (m = 1; 2; 3; …) that are fully transmitted. Using (10.8) we get

$$\lambda_m = \Delta s/m = 2d/m \cdot \sqrt{n^2 - \sin^2\alpha}.$$

If the incident monochromatic light is divergent a system of bright rings appears for the transmitted light, where all angles α give maximum transmission for which (10.8) is fulfilled.

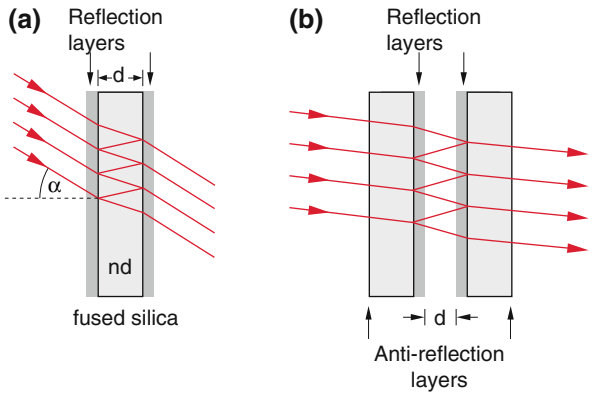


Fig. 10.21 Fabry-Perot interferometer. **a**) Etalon with reflecting coatings on both sides, **b**) two plates with reflecting surfaces on one side and anti-reflecting coatings on the other sides

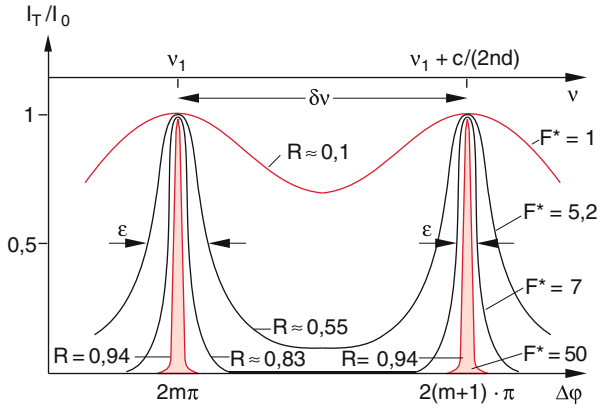


Fig. 10.22 Transmission $T = I_T/I_0$ of a plane-parallel plate for vertical incidence as a function of the phase difference of the interfering partial waves for different reflectivities

For illustration Fig. 10.22 shows the transmission $T(\Delta\varphi)$ for different values of the reflectivity R of each surface. This diagram shows that the transmission becomes $T = 1$ for $\Delta\varphi = 2m \cdot \pi$. This means that all incident light is transmitted. For $\Delta\varphi = (2m + 1)\pi$ the transmission becomes minimum and therefore the reflection maximum.

The full half width ε of the transmission peaks $I(\Delta\varphi)$ in Fig. 10.22, i.e. the phase difference $\varepsilon = \Delta\varphi_1 - \Delta\varphi_2$ with $I_T(\Delta\varphi_1) = I_T(\Delta\varphi_2) = I_0/2$ can be obtained from (10.25a) as

$$\varepsilon = 4 \arcsin\sqrt{(1/F)}$$

For sufficient small values of F (narrow transmission maxima of $I_T(\Delta\varphi)$ this becomes

$$\varepsilon = \frac{4}{\sqrt{F}} = \frac{2(1-R)}{\sqrt{R}}. \qquad (10.26a)$$

The full half width ε is the smaller the larger the reflectivity R is.

As has been mentioned before a plane parallel plate with thickness d and refractive index n acts as a spectral filter. For vertical incidence ($\alpha = 0$) the wavelengths $\lambda_m = 2nd/m$ have their maximum of transmission, while the wavelengths $\lambda_p = 4n \cdot d/(2p + 1)$ are maximal reflected.

The relative spectral half width of the transmitted intensity is with

$$\Delta\varphi = \varepsilon = \frac{2\pi\Delta s}{\lambda} - \frac{2\pi\Delta s}{\lambda + \Delta\lambda}$$
$$= 2\pi\Delta s \frac{\Delta\lambda}{\lambda \cdot (\lambda + \Delta\lambda)} \approx 2\pi \cdot m\lambda \frac{\Delta\lambda}{\lambda^2} \qquad (10.26b)$$
$$\frac{\Delta\lambda}{\lambda} = \frac{\varepsilon}{2\pi \cdot m} = \frac{1-R}{\pi \cdot m \cdot \sqrt{R}}$$

It depends on the reflectivity R of the surfaces and on the interference order m. With $\lambda = c/\nu \rightarrow d\lambda/d\nu = -(c/\nu^2)d\nu$ we obtain the relations

$$d\lambda/\lambda = -d\nu/\nu \qquad (10.26c)$$

Examples

1. $R = 0.55 \Rightarrow \varepsilon = 1.2 \approx 0.2 \cdot 2\pi \Rightarrow \Delta\lambda = 0.19 \cdot \lambda_m/m$. For $d = 1$ cm and $n = 1.5$, $\lambda = 500$ nm $\Rightarrow m = 6 \times 10^4 \Rightarrow \Delta\lambda/\lambda = 3.15 \times 10^{-6}$.
2. $R = 0.9 \Rightarrow \varepsilon = 0.21 \approx 0.03 \cdot 2\pi \Rightarrow \Delta\lambda = 0.03 \cdot \lambda_m/m \Rightarrow \Delta\lambda/\lambda_m = 5 \times 10^{-7}$.

This illustrates the large influence of the reflectivity on the half width of the transmission maxima.

10.4.1 Fabry-Perot-Interferometer

The multiple beam interference was already used in 1897 by the French Scientists Charles Fabry and Alfred Perot for the realization of high resolution interferometers, which have found increasing importance in modern optics and laser physics [6]. These **Fabry-Perot-Interferometers** (FPI) can be either a single plane parallel plate of optical glass or fused quartz with reflecting surfaces (Fig. 10.21a), or two plates which are coated only on one side. The coated sides of the two plates oppose each other and are carefully aligned to form a plane parallel air gap between them (Fig. 10.21b). In order to avoid reflections at the back sides of the plates they are either slightly wedge-shaped or the back sides are coated with an anti-reflection layer (see Sect. 10.4.3).

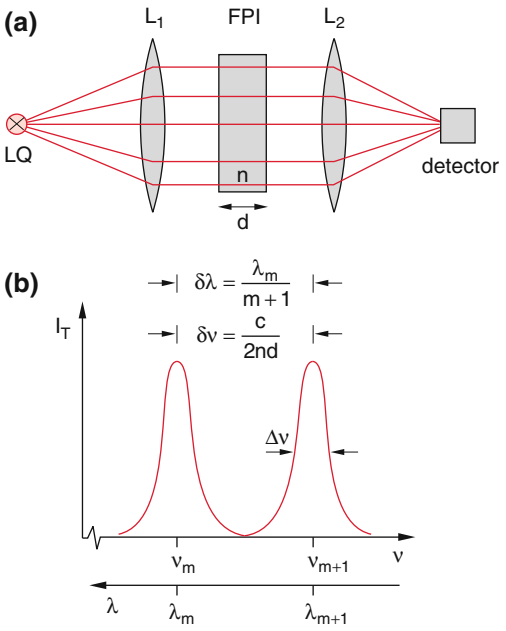


Fig. 10.23 Transmission of parallel incident light through a FPI. **a**) Experimental arrangement, **b**) transmitted intensity $I(\nu)$

We will illustrate the application of the FPI by two examples:

The light of a nearly point like source in the focal plane of a lens L_1 (Fig. 10.23) passes as parallel beam through the FPI and is imaged by a second lens L_2 onto the detector. The transmitted intensity $I_T(\Delta\varphi)$ depends for vertical incidence ($\alpha = 0$) on the phase difference

$$\Delta\varphi = (2\pi/\lambda) \cdot \Delta s = (4\pi/\lambda)n \cdot d$$

$$\Delta s = m \cdot \lambda_m \Rightarrow \lambda_m = \frac{2nd}{m} \tag{10.27}$$

In Fig. 10.23 the transmitted intensity $I(\lambda)$ resp. $I(\nu)$ is plotted. One sees that the function $I(\nu)$ is periodic with the period

$$\delta\nu = \nu_{m+1} - \nu_m = c/2nd \tag{10.28a}$$

In terms of the wavelength λ this becomes

$$\begin{aligned}\delta\lambda = \lambda_m - \lambda_{m+1} &= \frac{2nd}{m} - \frac{2nd}{m+1} \\ &= \frac{2nd}{m(m+1)} = \frac{\lambda_m}{m+1},\end{aligned} \tag{10.28b}$$

The distance $\delta\lambda$ resp. $\delta\nu$ between two successive transmission maxima is called the **free spectral range** of the interferometer.

The full half width $\Delta\nu = \nu_1 - \nu_2$ of the transmission maxima with the peak $I_T(\nu_m)$ is defined by

$$I_T(\nu_1) = I_T(\nu_2) = 1/2 \cdot I_T(\nu_m)$$

Inserting this into (10.25) we get

$$\Delta\nu = \frac{2}{\pi}\frac{\delta\nu}{\sqrt{F}} = \frac{c}{2nd}\frac{1-R}{\pi\sqrt{R}}. \tag{10.29}$$

The ratio of free spectral range $\delta\nu$ to the full half width $\Delta\nu$

$$F^* = \frac{\delta\nu}{\Delta\nu} = \frac{\pi \cdot \sqrt{R}}{1-R} \tag{10.30}$$

is called the **finesse of the interferometer**. It is a measure of the number of interfering transmitted or reflected partial waves.

This can be seen as follows:

The width of the transmission maxima is determined by the number p of interfering partial waves. If Δs is the path difference between consecutive interfering beams then the free spectral range is

$$\Delta\nu = \frac{c}{\Delta s}.$$

The path difference between the first and the p-th beam is then $p \cdot \Delta s$ and the half width of the maxima is

$$\Delta\nu = \frac{c}{p \cdot \Delta s}$$

The finesse F^* is then

$$F^* = \frac{\delta\nu}{\Delta\nu} = p.$$

The half width $\Delta\nu = \delta\nu/F^*$ of the transmission maxima of the interferometer is the ratio of free spectral range and finesse F^*.

Example

$R = 0.98 \Rightarrow F^* \approx 155$. This means 155 partial waves interfere with each other. With an optical thickness $n \cdot d = 3$ cm $\Rightarrow \delta\nu = c/(2n \cdot d) = 5 \times 10^9$ s$^{-1} \Rightarrow \Delta\nu = \delta\nu/F^* = 3.2 \times 10^7$ s$^{-1} = 32$ MHz.

Remark The considerations above have anticipated that the reflecting surfaces are ideal planes, which are aligned exactly plane parallel. In reality the surfaces deviates from ideal planes and show small surface irregularities and micro roughness. With the maximum distortion $2\pi/q$ of the phase front of a wave after reflection by the surface the flatness of the surface is defined as λ/q. After p transits with p reflections the maximum deviation of the phase front from an ideal plane is

$$\Delta\varphi = (p/q) \cdot 2\pi$$

For $p = q/2$ the phase difference between the first and the pth partial wave has increased to $\Delta\varphi = \pi$. This partial wave experiences therefore destructive instead of constructive interference which diminishes the reflected intensity.

Furthermore any misalignment that causes a deviation from the exact plane-parallel position of the two reflecting surfaces causes a variation of the phase difference across the beam diameter. This results in a decrease of the number p of transits with constructive interference and causes a reduction of the finesse F^*. Also diffraction effects at the edges of the beam result in a deviation of the phase front from an ideal plane and cause a reduction of the finesse.

The total finesse F_t^* of a Fabry-Perot interferometer is therefore smaller than the reflectivity finesse F^* defined in (10.30). It is defined as

$$\frac{1}{F_g^*} = \sqrt{\sum_i \left(\frac{1}{F_i^*}\right)^2}, \tag{10.31}$$

where the different summands in the square root describe the influence of surface defects, misalignment and diffraction on the spectral width of the transmission maxima. It is therefore useless to increase the reflectivity above a limit which is set by all other effects that decrease the total finesse. It turns out that the optimum choice of the reflectivity R is reached for $p = q$.

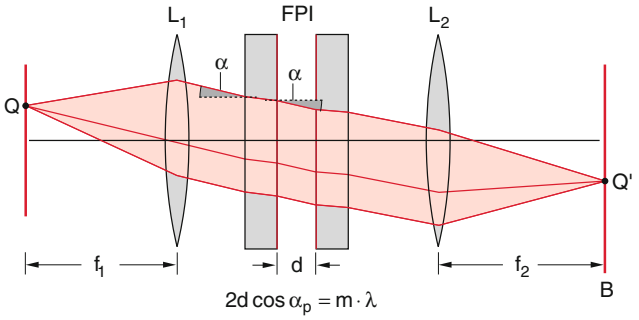


Fig. 10.24 Formation of a ring system behind the FPI under illumination by an extended light source

Generally the illumination of the FPI is not realized by a point-like source but by an extended light source. We regard in Fig. 10.24 a FPI with a plane parallel air gap (refractive index $n \approx 1$) which is illuminated by an extended light source LS in the focal plane $z = z_0 = f_1$ of the lens L_1 perpendicular to the symmetry axis z. Light emitted by a point Q of the light source passes through the FPI as parallel light beam under the angle α against the z-axis. Only for those angles $\alpha_p (p = 1, 2, 3\ldots)$ maximum transmission occurs which fulfill the condition

$$\begin{aligned} \Delta s &= 2d\sqrt{n^2 - \sin^2 \alpha_p} = 2d \cos \alpha_p \\ &= m \cdot \lambda \end{aligned} \tag{10.32}$$

with m = integer. For a monochromatic light source the transmitted light therefore shows a pattern of concentric bright rings (Fig. 10.25). The acuity of the rings depends on the total finesse F_t^* of the FPI.

When the parallel light passing through the FPI is imaged by a second lens L_2 with focal length f_2 onto the observation plane B the diameters D_p of the rings are

$$D_p = 2f_2 \cdot \tan \alpha_p \approx 2f_2 \cdot \alpha_p. \tag{10.33}$$

The diameter $D = D_0$ of the smallest ring obeys the condition (10.32) for $m = m_0$. For this ring is $\alpha \ll 1$ and we

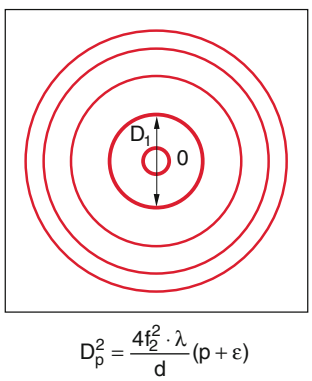


Fig. 10.25 Ring system of the transmitted light through a plane FPI emitted by an extended light source

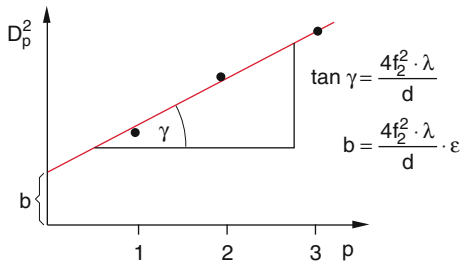


Fig. 10.26 Determination of the wavelength λ from the slope and the intercept of the straight line $D_p^2(p)$, plotting the square of the ring diameters against the ring number p

can approximate $\cos \alpha \approx 1 - \alpha^2/2$. We then obtain the condition (10.23) for constructive interference

$$\begin{aligned} &2d(1 - \alpha_p^2/2) = m_p \cdot \lambda = (m_0 - p)\lambda \\ &\Rightarrow 2d = \left(m_0 + \frac{d\alpha_0^2}{\lambda}\right)\lambda = (m_0 + \varepsilon)\lambda. \end{aligned} \tag{10.34}$$

the quantity $\varepsilon = d\alpha_0^2/\lambda < 1$ is the **excess** of the FPI.

For $\alpha_0 = 0$ is $\varepsilon = 0$. In this case an integer number of half wavelengths fits between the two parallel reflecting surfaces, i.e. $m_0 \cdot \lambda/2 = n \cdot d$.

For $\alpha_0 \neq 0$ the quantity ε gives the excess $\varepsilon = d/(\lambda/2) - m_0$ of the optical distance d between the two surfaces in units of $\lambda/2$ over the integer m_0. This means that the rings appear for plate separations that are non-integer multiples of $\lambda/2$.

For the squares D_p^2 of the ring diameters D_p we get from (10.34)

$$D_p^2 = \frac{4f_2^2 \cdot \lambda}{d}(p + \varepsilon). \tag{10.35}$$

Plotting the squared ring diameters D_p^2 against the ring number p allows the determination of λ from the slope of the straight line $D_p^2(p)$, if the plate separation d is known (Fig. 10.26). The intercept with the axis $p = 0$ gives the excess ε. The distance d can be obtained when the ring diameters are measured with a known calibration wavelength λ_c.

10.4.2 Dielectric Mirrors

With metal mirrors (surfaces covered with metal layers such as aluminum, silver or gold) the maximum reflectivity is $R = 0.95$. In reality generally only $R = 0.90$ is reached. The reason for this upper limit is the high absorbance of metals in the visible range. The reflectivity is mainly determined by the imaginary part of the complex refractive index (see Sect. 8.4.9).

This limited reflectivity of metal mirrors is for many applications not sufficient. For example laser mirrors generally demand a reflectivity of $R > 0.98$ (see Vol. 3). A higher reflectivity up to $R = 0.999$ can be achieved by

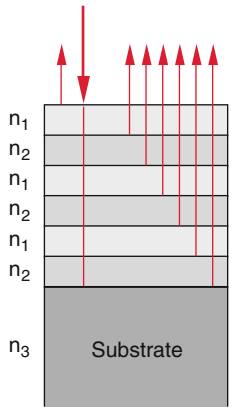


Fig. 10.27 Dielectric mirror with glass substrate and many thin dielectric layers with small absorption and alternating different refractive indices n_1 and n_2. The thickness of the layers is exaggerated. It is in reality only about 1 µm

constructive interference between the partial waves reflected by many thin dielectric layers with different refractive indices, but small absorbance (Fig. 10.27). In order to achieve maximum reflectance R all partial waves, reflected by the different layers must be in phase. This will be illustrated by the example of vertical incidence ($\alpha = 0$) onto a dielectric mirror with two layers and zero absorbance (Fig. 10.28). For this case a phase jump $\Delta\varphi = \pi$ occurs under reflection at the optical thicker medium, whereas no phase jump appears for the reflection at the optically thinner medium.

If the refractive indices follow the sequence $n_{\text{air}} < n_1 > n_2 > n_3$ the incident light suffers a phase jump only for the reflection at the upper surface.

Constructive interference is obtained if the optical thickness of the upper layer is $n_1 \cdot d_1 = \lambda/4$ and $n_2 \cdot d_2 = \lambda(2)$.

The reflectivity of the three interfaces are

$$R_1 = \left(\frac{n_1 - 1}{n_1 + 1}\right)^2, \quad R_2 = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2, \quad R_3 = \left(\frac{n_2 - n_3}{n_2 + n_3}\right)^2. \tag{10.36a}$$

The total reflected amplitude is then

$$\begin{aligned} A_R &= A_1 + A_2 + A_3 \\ &= A_0\sqrt{R_1} + (A_0 - A_1)\sqrt{R_2} \\ &\quad + (A_0 - A_1 - A_2)\sqrt{R_3}. \end{aligned} \tag{10.36b}$$

The reflected intensity is with the amplitudes

$$\begin{aligned} A_2 &= A_0(\sqrt{R_2} - \sqrt{R_1 R_2}) \\ A_3 &= A_0(\sqrt{R_2} - \sqrt{R_1 R_3} - \sqrt{R_2 R_3} + \sqrt{R_1 R_2 R_3}) \end{aligned} \tag{10.36c}$$

$$\begin{aligned} I_R &= \varepsilon_0 \cdot c\left|\sum_{p=1}^{3} A_p\right|^2 \\ &= \varepsilon_0 \cdot cA_0^2\left[\sqrt{R_1}(1 - \sqrt{R_2} - \sqrt{R_3})\right. \\ &\quad \left. + \sqrt{R_2}(2 - \sqrt{R_3} - \sqrt{R_1\sqrt{R_3}})\right]^2 \end{aligned} \tag{10.36d}$$

and the reflectivity is $R = I_R/(\varepsilon_0 \cdot cA_0^2)$

Nowadays it is possible the reach values of $R > 0.999$ for a selected wavelength λ with 15–20 layers [7]. Figure 10.29 shows the reflection curve $R(\lambda)$ of a dielectric mirror with 12 layers.

As material for layers with a low refractive index $MgF_2(n = 1.38)$ or $SiO_2(n = 1.46)$ are, for instance, chosen, while the layers with high values of n consist, for example, of titanium oxide $TiO_2(n = 2.4)$.

For multilayer dielectric mirrors the amplitudes of the reflected partial waves are

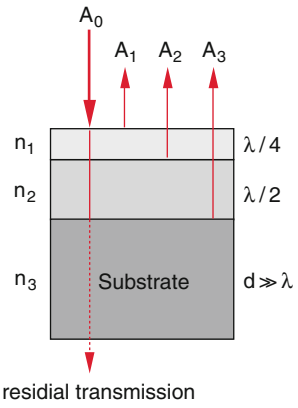


Fig. 10.28 Superposition of the reflected partial waves for a dielectric two-layer mirror with $n_1 > n_2 > n_3$

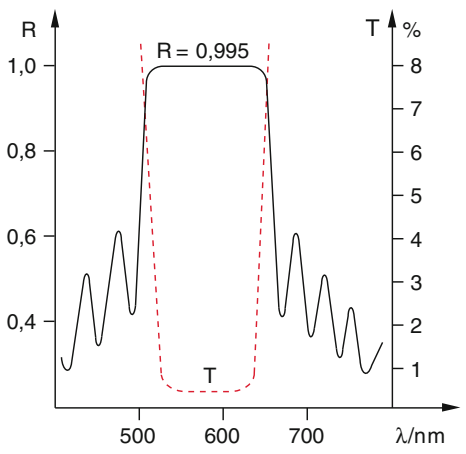


Fig. 10.29 Reflectivity $R(\lambda)$ of a dielectric multilayer mirror

$$
\begin{aligned}
|A_1| &= \sqrt{R_1}|A_0|;\\
|A_2| &= (1-R_1)\sqrt{R_2}|A_0|;\\
|A_3| &= (1-R_1)\sqrt{R_2}|A_0|;\\
|A_4| &= (1-R_1)R_2^{3/2}R_1|A_0|;\\
|A_5| &= (1-R_1)R_2^2R_2^{3/2}|A_0| \quad \text{etc.}
\end{aligned}
\tag{10.36e}
$$

The intensities are proportional to the squares of the amplitudes.

The exact calculation of the reflectivity $R(\lambda) = I_R(\lambda)/I_0$ of multilayer dielectric mirrors and the selection of the different layers demands sophisticated computer programs [7, 8].

10.4.3 Anti-reflection Coating

In order to minimize the often disturbing and unwanted reflections at glass surfaces (for instance at eye glasses or at the lenses of a camera objective) the surfaces of the lenses are covered with a thin layer that suppresses reflection by destructive interference (Fig. 10.30a). The phase difference between the two partial waves reflected by the two interfaces must be $\Delta\varphi = (2m + 1) \cdot \pi$. We will restrict the discussion to vertical incidence ($\alpha = 0$).

Part of the incident wave is reflected at the interface air ($n = 1$) − antireflection layer ($n_1 > 1$). It suffers a phase jump of π. The transmitted part is partly reflected at the second interface layer-glass.

The amplitudes of the reflected partial waves for a two-layer antireflection coating (Fig. 10.30b) can be calculated similar to (10.36e). The difference is only that here destructive instead of constructive interference is optimized. One obtains for the amplitudes of the partial waves reflected by the 4 interfaces

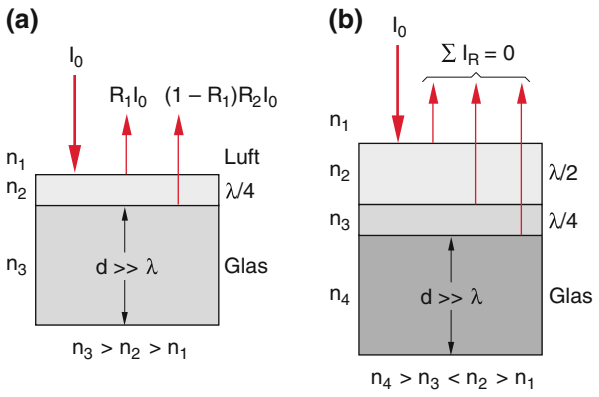


Fig. 10.30 Anti-reflection coating. **a**) Single layer, **b**) two layers

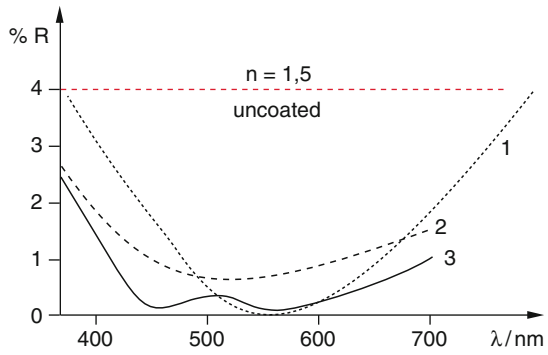


Fig. 10.31 Residual reflectivity $R(\lambda)$ for a single anti-reflection layer (curve 1) compared with a an uncoated glass surface with $n = 1.5$. The curve 2 represents a two-layer broad band anti-reflection coating, curve 3 a three layer coating

$$
\begin{aligned}
|A_1| &= \sqrt{R_1}|A_0|\ ;\\
|A_2| &= (1-R_1)\sqrt{R_2}|A_0|\ ;\\
|A_3| &= (1-R_1)R_2\sqrt{R_1}|A_0|\ ;\\
|A_4| &= (1-R_1)R_2^{3/2}R_1|A_0|\ ;\\
|A_5| &= (1-R_1)R_2^2R_1^{3/2}|A_0| \quad \text{etc.}
\end{aligned}
\tag{10.36f}
$$

For a one-layer antireflection coating only three reflected partial waves must be considered. Complete extinction of the total reflected amplitude is obtained if

$$
\begin{aligned}
n_2 &= \sqrt{n_{\text{Luft}} \cdot n_3}\\
&\approx 1.225 \quad \text{for} \quad n_3 = 1.5
\end{aligned}
\tag{10.37a}
$$

The optical thickness of the layer is then (see Problem 10.12)

$$
d = \frac{(2m+1)}{4}\frac{\lambda_0}{n_2} \quad \text{with} \quad m = 0, 1, 2, \ldots \tag{10.37b}
$$

For the single layer a complete suppression of the reflected wave ($I_R(\lambda_0) = 0$) can be obtained only for a selected wavelength λ_0 (Fig. 10.31). Using several layer the reflection can be minimized for a broader spectral range. With a two-layer antireflection coating for instance the residual reflected intensity is below 1% over the whole visible range, compared with 4% for an uncoated glass surface.

10.4.4 Applications of Interferometers

An important technical application field is the accurate measurement of distances and lengths. Since here phase differences between the partial waves reflected by the two ends of a measured distance can be determined, the uncertainty of the length measurement is only a small fraction of the visible wavelength. One example is the Michelson interferometer in

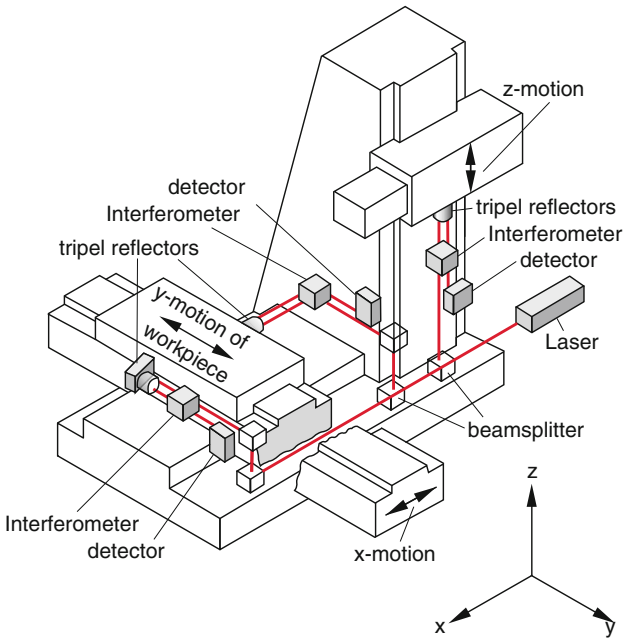


Fig. 10.32 Principle of an interferometric controlled machine tool movable in three directions. The work piece is moved within the *x*-*y*-plane. The tool is moved in the *z*-direction

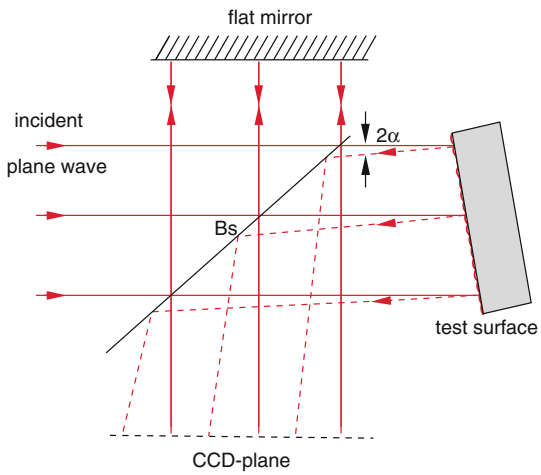


Fig. 10.33 Measurement of small deviations from a nearly plane surface, using a Michelson interferometer. In the plane of the CCD-detector an interference stripe pattern occurs which is deformed by any deviation of the test surface from an ideal plane

Fig. 10.10. When the mirror M_2 moves by the length $\Delta s = (m+\varepsilon)\cdot\lambda$ with $\varepsilon<1$ one observes $2(m+\varepsilon)$ interference maxima in the observation plane B. For a sufficient high signal-to-noise ratio the excess ε can be determined by interpolation within $0.01\cdot\varepsilon$. This allows the determination of the length within $\lambda/100$. This interferometric measurement is utilized for the control of the movement of machine tools in three dimensions (Fig. 10.32) where three interferometer arms are necessary for the very accurate control of the three-dimensional motion of the machine with an accuracy of better than 50 nm.

This precision is necessary for the production of wavers for integrated circuits. Since the waver has to be precisely positioned at the same position within 50 nm after each process step, this can be only achieved with interferometric methods.

Another example is the inspection of surfaces in order to determine local deviations from the ideal surface. The measurement principle illustrated in Fig. 10.33 uses a Michelson interferometer. The inspected surface is illuminated by an expanded laser beam. The light reflected by the slightly tilted surface is superimposed at the beam splitter *BS* with a reference beam in the second arm of the interferometer. An ideal plane would produce straight parallel stripes in the observation plane B with a distance $\Delta = \lambda/2\cdot\sin(2\alpha)$ where α is the tilt angle. Any deviation from the ideal plane appears as distortion of the interference stripes. The degree of distortion is a measure of the geometrical magnitude of the deviation and the number of the distorted stripe gives information about the location of the deviation on the surface.

This method is applied for the inspection of extremely flat or spherical mirrors where deviations of less than $\lambda/100$ are demanded.

Another example is the measurements of local variations of the refractive index $n(x, y)$, which can be determined with a Mach-Zehnder interferometer (Fig. 10.34a). Such local variations in air are for instance produced by the temperature gradient above a candle flame. The interference stripes are deformed in this region (Fig. 10.34b). The degree of deformation allow the determination of the local temperature profile. For further examples see [9].

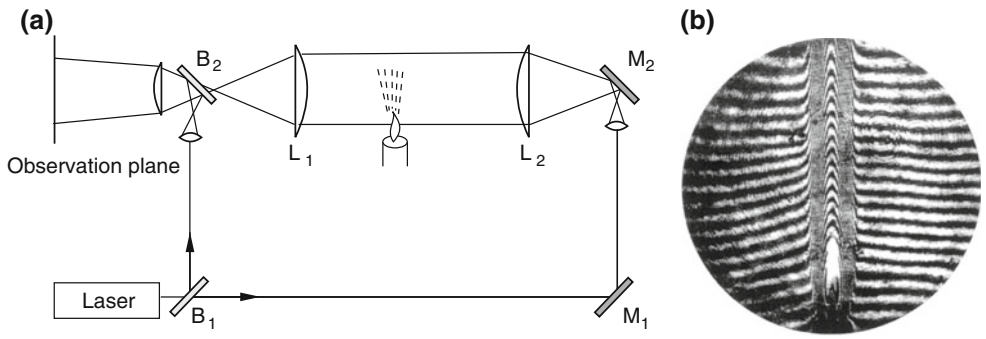


Fig. 10.34 Measurement of local variations of the refractive index n. **a**) Mach Zehnder interferometer, **b**) variation of $n(x, y)$ above a candle flame visible by the distortion of the interference fringes

10.5 Diffraction

When a light beam passes through apertures or along edges of transparent media part of the light is diffracted out of its original propagation direction. Light is then observed in directions that are not allowed in geometrical optics. This phenomenon is called **diffraction**.

10.5.1 Diffraction as Interference Phenomenon

We regard in Fig. 10.35 N oscillators that are regularly located with the mutual distance d on the x-axis. They are induced to oscillations by a plane wave propagating into the z-direction. These induced oscillators radiate secondary waves which are in the plane $z = z_0$ all in phase. When calculating the total amplitude of all waves emitted into the direction θ by the N oscillators on the x-axis we have to take into account the different path lengths for the partial waves which differ by $\Delta s = d \cdot \sin\theta$ between waves from adjacent oscillators. If the amplitudes of all partial waves are equal ($A_i = A$) we get the phase difference

$$\Delta\varphi = \frac{2\pi}{\lambda}\Delta s = \frac{2\pi}{\lambda} d \cdot \sin\theta \tag{10.38a}$$

where we have set the phase of the first wave $\varphi_1 = 0$. The sum of the different contributions assuming the same amplitude gives the total amplitude is

$$E = A.\sum_{j=1}^{N} e^{i(\omega t + \varphi j)} = A \cdot e^{i\omega t}\sum_{j=1}^{N} e^{i\cdot(j-1)\Delta\varphi}, \tag{10.38b}$$

The sum of the geometrical series is

$$\sum_{j=1}^{N} e^{i(j-1)\Delta\varphi} = \frac{e^{iN\Delta\varphi} - 1}{e^{i\Delta\varphi} - 1}$$
$$= e^{i\frac{N-1}{2}\Delta\varphi} \cdot \frac{e^{i\frac{N}{2}\Delta\varphi} - e^{-i\frac{N}{2}\Delta\varphi} - 1}{e^{i\Delta\varphi/2} - e^{-i\Delta\varphi/2}} \tag{10.39}$$
$$= e^{i\frac{N-1}{2}\Delta\varphi} \cdot \frac{\sin[(N/2)\Delta\varphi]}{\sin(\Delta\varphi/2)}.$$

The intensity $I = c\varepsilon_0|E|^2$ of the total wave in the direction θ is then with (10.38a)

$$I(\theta) = I_0 \cdot \frac{\sin^2[N\pi(d/\pi)\sin\theta]}{\sin^2[\pi(d/\pi)\sin\theta]}, \tag{10.40}$$

The shape of the function $I(\theta)$ depends strongly on the ratio d/λ .

For $d < \lambda$ the function $I(\theta)$ has only one maximum for $\theta = 0$ and decreases to zero for increasing values of θ (see Vol. 1, Sect. 11.11). For small values of θ we can approximate $\sin\theta \approx \theta$. For $d < \lambda$ and $\sin\theta \ll 1$ is also $\pi(d/\lambda)\sin\theta \ll 1$ and we can reduce (10.40) to

$$I(\theta) = N^2 I_0 \cdot \frac{\sin^2 x}{x^2} \tag{10.41}$$

with $x = N\pi(d/\lambda)\sin\theta$. The function $(\sin x/x)^2$ is plotted in Fig. 10.36. This illustrates that the function has higher values only in the range $-\pi < x < +\pi$ ($\lambda/(N \cdot d) < \sin\theta < +/(N \cdot d)$) of the central maximum. The area of this central intensity maximum

$$\int_{-\pi}^{+\pi} \frac{\sin^2 x}{x^2} dx \approx 0.9 \cdot \int_{-\infty}^{+\infty} \frac{\sin^2 x}{x^2} dx \tag{10.42}$$

contains already about 90% of the total intensity diffracted into all directions θ.

When the width $D = N \cdot d$ of the oscillator arrangement is large compared to the wavelength ($D \gg \lambda$) it follows for the range $|x| < \pi$ the angular range $|\sin\theta| \ll x/\pi < 1$. This means that the diffracted intensity has noticeable values only in a very narrow angular range $|\Delta\theta| = \lambda/(N \cdot d) << 1$ around $\theta = 0$.

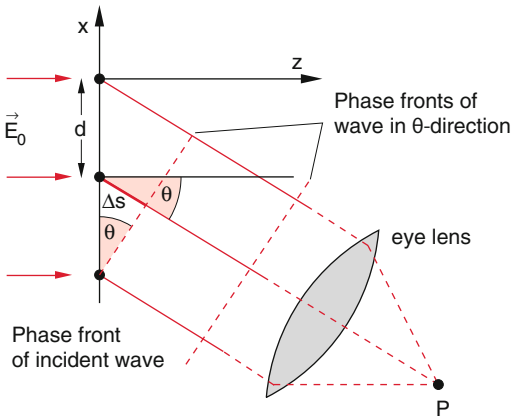


Fig. 10.35 Illustration of Eq. (10.40)

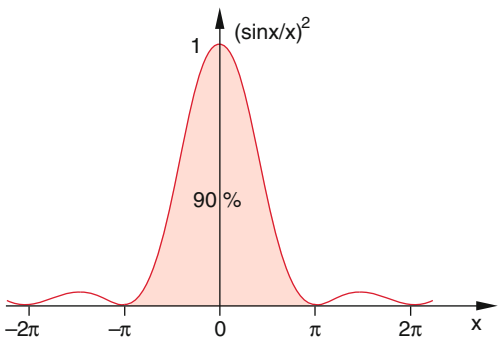


Fig. 10.36 The function $\sin^2 x/x^2$

Example

$D = 1$ cm, $\lambda = 500$ nm $\Rightarrow \sin\theta < 5 \times 10^{-7}/10^{-2} = 5 \times 10^{-5}$. The central intensity maximum has an angular width of $\Delta\theta < 0.003°$.

This result illustrates the following astonishing fact: Although each single oscillator radiates its intensity into all directions $-\pi < \theta < +\pi$, the superposition of the radiation from many regularly spaced oscillators with a distance $d < \lambda$ leads to a total radiation intensity that is essentially emitted in forward direction within the narrow angular range $\theta = 0 \pm \Delta\theta$ with $\Delta\theta = \lambda/(N \cdot d) \ll 1$ which depends on the total width $N \cdot d$ of the oscillator arrangement.

The angular width of the central intensity maximum between the maximum of $\sin^2 x/x^2$ at $x = 0$ and the zero points at $x = \pi \Rightarrow \Delta\theta = \lambda/(N \cdot d) = \lambda/D$.

For $D \to \infty \Rightarrow \Delta\theta \to 0$ (Fig. 10.37).

The propagation of the waves in directions $\theta \neq 0$ is called **diffraction**. We see from the above considerations that diffraction can be explained by interference of many partial waves. It comes about because of the finite extension of the oscillators or the limiting cross section of the incident wave.

Note The coherent superposition of N equal amplitudes A_i of the partial waves emitted by N oscillators is $A_t = \Sigma A_i = N \cdot A$. The total intensity is, however, $I_t = c \cdot \varepsilon_0 \cdot (\Sigma A_i)^2 = c \cdot \varepsilon_0 \quad N^2 \cdot A^2 = N^2 \cdot I_0$. This means that N oscillators, coherently excited, do not have the total intensity $N \cdot I_0$ (as one might suggest at first sight) but $N^2 \cdot I_0$.

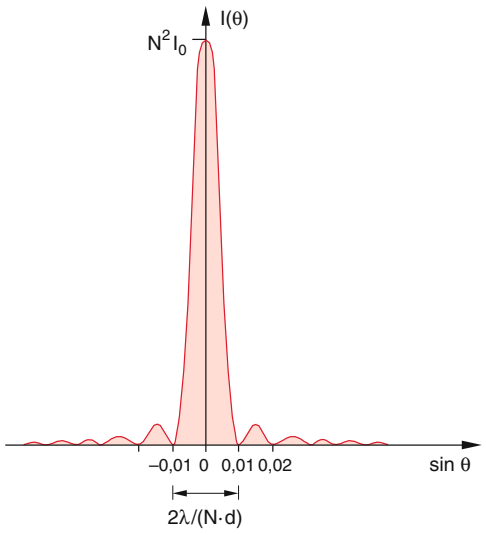


Fig. 10.37 The scattered intensity $I(\theta)$ for $d < \lambda$ and $D = N \cdot d = 100\,\lambda$. The width between the two closest zero point on both sides of $I(\theta)$ is $\Delta\theta = 2\lambda/(N \cdot d)$

10.5.2 Diffraction by a Slit

When we apply the above considerations to the propagation of a plane wave through a slit with width b we have to take into account the following:

Each point within the slit is a source of a spherical wave, because the electric and the magnetic field of the incident wave change with time in P and are therefore according to the Maxwell equations even in vacuum the origin of new electromagnetic waves. These secondary waves superimpose (Huygens principle see Vol. 1, Sect. 11.11).

If we replace each oscillator by a line segment Δb (Fig. 10.38) with continuously distributed emitters, the slit contains $N = b/\Delta b$ emitting line segments. Their amplitude is $A = N \cdot A_0 \cdot \Delta b/b$. Instead of (10.40) we then obtain

$$I(\theta) = N^2 I_0 \left(\frac{\Delta b^2}{b}\right) \frac{\sin^2[\pi(b/\pi)\sin\theta]}{\sin^2[\pi(\Delta b/\pi)\sin\theta]}, \qquad (10.43)$$

where I_0 is the intensity emitted by one emitter line segment. With the abbreviation $x = \pi \cdot (b/\lambda) \cdot \sin\theta$ and $\Delta b = b/N$ this becomes

$$I(\theta) = I_0 \cdot \frac{\sin^2 x}{\sin^2(x/N)}. \qquad (10.44)$$

Now we consider the limit $N \to \infty$, i.e. $\Delta b \to 0 \Rightarrow \sin^2(x/N) \to x^2/N^2$. This implies a continuous spatial intensity distribution. The total intensity, transmitted by the slit is $I_s = N^2 I_0$, which can be written as

$$\lim_{N\to\infty} I(\theta) = N^2 I_0 \cdot \frac{\sin^2 x}{x^2} = I_S \cdot \frac{\sin^2 x}{x^2}. \qquad (10.45)$$

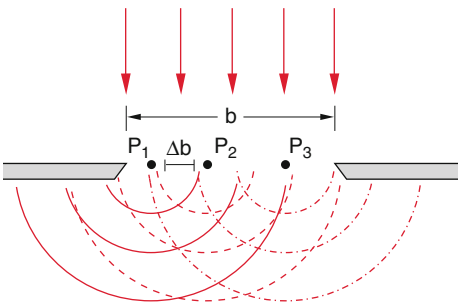


Fig. 10.38 Diffraction at a slit

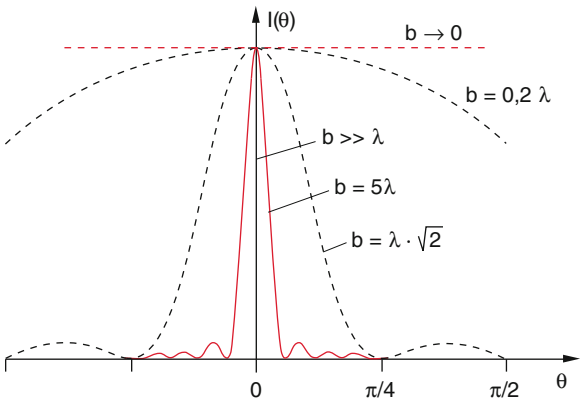


Fig. 10.39 Angular intensity distribution of light diffracted by a slit for different ratios λ/b of wavelength λ to slitwidth b

because $\int(\sin^2 x/x^2)\mathrm{d}x = 1$.

This function, which has been shown already in Fig. 10.36, is again illustrated in Fig. 10.39 as a function of the diffraction angle θ. Most of the light propagates straight on ($\theta = 0$). The intensity $I(\theta)$ becomes zero for $\sin\theta = \lambda/b$ but has many more maxima for $\sin\theta > \lambda/b$ which become smaller and smaller with increasing θ. This can be vividly understood by Fig. 10.40. For $\sin\theta = \lambda/b$ the path difference between the edges of the transmitted light beam is just $\Delta s_m = \lambda$. We can divide the beam into two halves. To each partial light beam in the first half exists a partial beam in the second half with a path difference of $\lambda/2$. All these corresponding partial waves therefore suffer destructive interference and cancel each other. Therefore zero intensity appears for $\Delta s_\mathrm{m} = d \cdot \sin\theta = \lambda$.

For $\Delta s_\mathrm{m} = (3/2)\lambda$ we divide the total wave into three sections. Two of these sections cancel each other while the third section is left over. This corresponds to the first side maximum of $I(\theta)$. The central maximum contains 90% of the total transmitted intensity, independent of the ratio λ/b (see 10.42).

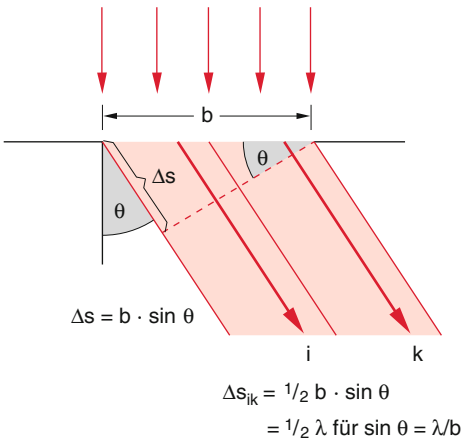


Fig. 10.40 Illustration of the intensity minimum for $\sin\theta = \lambda/b$

In Fig. 10.39 the intensity distribution $I(\theta)$ is plotted for different ratios λ/b. For $b \gg \lambda$ the central maximum of $I(\theta)$ becomes very narrow, i.e., the angular full width $\Delta\theta = 2\lambda/b$ between the two zero points becomes small. The transmitted light propagates essentially straight on.

Example

$$b = 1000\lambda \Rightarrow \Delta\theta = 2 \times 10^{-3}\,\mathrm{rad} = 0.11°.$$

Note, however, that in spite of this small diffraction angle the light beam is slightly divergent and its diameter increases with increasing distance from the slit. For λ = 500 nm and a slit width of b = 0.5 mm at a distance d behind the slit the beam diameter has increased to $b + d \cdot \Delta\theta$. This gives for d = 10 m and $\Delta\theta = 2 \times 10^{-3}$ rad a diameter of 20.5 mm. The diffraction has caused an increase of the beam diameter by a factor of 41!

For $b \leq \lambda$ no minimum of $I(\theta)$ exists because it should occur for $\sin\theta = \lambda/b$ and $|\sin\theta| \leq 1$ cannot be larger than 1. In this case the central maximum is spread out over the whole half space behind the slit. Therefore one cannot see any diffraction pattern, but a monotonically decreasing intensity $I(\theta)$ in the angular range $0 \leq \theta \leq \pi/2$.

The angular intensity distribution $I(\theta)$ of a monochromatic wave with wavelength λ, transmitted through a slit with width b depends on the ratio λ/b. For $\lambda/b \ll 1$ a central diffraction maximum appears with an angular width $\Delta\theta = 2\lambda/b$ between the two zero points and furthermore small side maxima at $\theta_\mathrm{m} = \pm(2m+1)\lambda/2b$. For $\lambda/b > 1$ the intensity of the central maximum is spread out over the whole angular range $|\theta| \leq 90°$.

When passing through a circular aperture with radius R the diffracted intensity distribution $I(\theta)$ has to show rotational symmetry around the symmetry axis of the aperture (Fig. 10.41). The more complex calculation [10] gives instead of (10.43) the distribution

$$I(\theta) = I_0 \cdot \left(\frac{2J_1(x)}{x}\right)^2 \qquad (10.46)$$

with

$$x = \frac{2\pi R}{\lambda} \cdot \sin\theta,$$

where $J_1(x)$ is the first order Bessel-function. The distribution (1.46) has zero points at $x_1 = 1.22\pi$, $x_2 = 2.16\pi, \ldots$, The first zero point of $I(\theta)$ therefore appears at $\sin\theta_1 = 0.61\lambda/R$. The position of the side maxima and their intensities are:

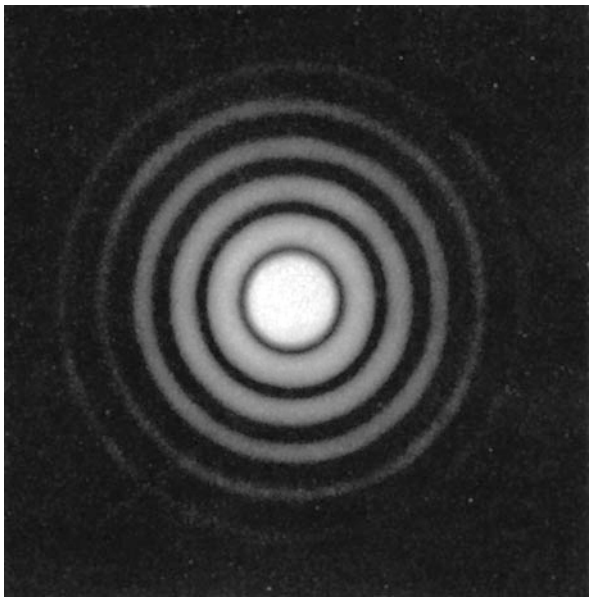


Fig. 10.41 Diffracted ring system observed when a parallel light passes through a circular aperture. (from: M. Cagnac, M. Francon, J.C. Thrier: Atlas optischer Erscheinungen Springer Berlin Göttingen 1962)

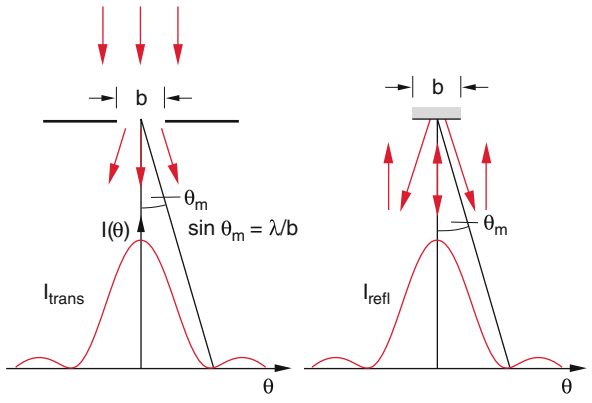


Fig. 10.42 Equivalent diffraction patterns for the diffraction at an aperture or a mirror with equal width b

$$
\begin{aligned}
I_{M_1} &= 0.0175 I_0 \quad \text{at} \quad \sin\theta_{M_1} = 0.815\lambda/R, \\
I_{M_2} &= 0.00415 I_0 \quad \text{at} \quad \sin\theta_{M_2} = 1.32\lambda/R, \\
I_{M_3} &= 0.0016 I_0 \quad \text{at} \quad \sin\theta_{M_3} = 1.85\lambda/R.
\end{aligned}
$$

Note Diffraction phenomena not only appear when a light beam passes through a limiting aperture, but also when a light wave is reflected by a mirror with limited cross section (Fig. 10.42). The intensity pattern of light reflected by a circular mirror with diameter $2R$ equals exactly that of light passing through a circular aperture with radius R (see Sect. 10.7.5).

10.5.3 Diffraction Gratings

When a plane wave incides onto an arrangement of N parallel slits in the plane $z = 0$ (diffraction grating Fig. 10.43) the intensity distribution $I(\theta)$ is determined by two factors:

- The interference between the light beams through the different slits. This distribution corresponds exactly to the coherent emission of N oscillators treated in Sect. 10.5.1, resulting in the intensity distribution (10.40).
- the intensity distribution (10.43), due to the diffraction by each slit.

With the slit width b and the distance d between adjacent slits we get, according to (10.43) and (10.40) the angular intensity distribution

$$
I(\theta) = I_S \cdot \frac{\sin^2[\pi(b/\lambda)\sin\theta]}{[\pi(b/\lambda)\sin\theta]^2} \cdot \frac{\sin^2[N\pi(d/\lambda)\sin\theta]}{\sin^2[\pi(d/\lambda)\sin\theta]^2}, \tag{10.47}
$$

where θ is the angle against the z-axis and I_s is the intensity transmitted by each slit. The first factor describes the diffraction by a single slit and the second factor the interference between the light transmitted by N slits.

Maxima of $I(\theta)$ appear for those directions θ for which the path difference between adjacent slits

$$
\Delta s = d \cdot \sin\theta = m \cdot \lambda \tag{10.48}
$$

becomes an integer multiple m of the wavelength λ. The intensity of these maxima depends on the diffraction distribution of the single slits, i.e. on the first factor in (10.48). The diffraction ensures that altogether the transmitted intensity reaches the angular range $\theta > 0$. The larger the slit

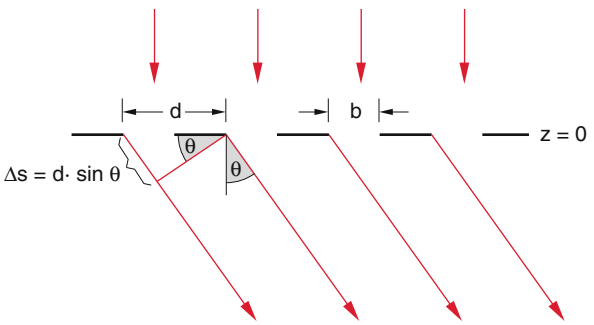


Fig. 10.43 Diffraction grating with N parallel slits, which is illuminated perpendicular by a plane wave

width b is, the smaller is the angular range θ which can be reached by the diffracted light.

In Fig. 10.44 is as example the distribution $I(\theta)$ for a diffraction grating with 8 slits and a ratio $d/b = 2$ illustrated.

The different maxima are called *diffraction maxima of mth-order* (a better name would be i*nterference* maxima). As can be seen from (10.48) the maximum order is (because of $\sin\theta < 1$)

$$m_{\max} = d/\lambda,$$

It is determined by the ratio of slit distance d and wavelength λ. The principal maxima occur, when the denominator of the second factor in (10.47) becomes zero (the nominator becomes then also zero and the value of the fraction can be obtained by the rule of de l'Hopital, which gives N^2 for the second factor in (10.47). The intensity of the principal maxima is determined by the diffraction distribution described by the first factor (dashed curve in Fig. 10.44).

Between the principal maxima $N-2$ side maxima occur at such angles θ_p, for which the nominator of the second factor becomes 1 and the denominator is $\neq 0$. On gets

$$\sin\theta_p = \frac{(2p+1)\lambda}{2N \cdot d} \quad (p = 1, 2, \ldots, N-2).$$

The magnitude $I(\theta_p)$ of these side maxima can be obtained from the second factor in (10.47). The result is

$$I(\theta_p) = \frac{I_0}{N^2} \frac{1}{\sin^2[(2p+1)\pi/(2N)]},$$

For odd values of N the intensity of the mid side maximum at $(p = (N-1)/2)$ becomes $I = I_0/N^2$. For sufficiently large values of N these side maxima are therefore negligible. For example, in modern optical diffraction gratings is $N = 10^5$ which gives

$$I = 10^{-10} \cdot I_0.$$

Figure 10.44 illustrates that the intensity of the principal diffraction maxima of mth order depends on the angular width of the diffraction intensity. The slit width b must be therefore sufficiently small in order to diffract sufficient intensity at least into the first order interference maximum.

Diffraction gratings play an important role in spectroscopy for the measurement of optical wavelengths. For a sufficiently high spectral resolution one needs gratings with $N = 10^5$. For a grating with total width $D = 10$ cm this implies a slit distance of 1 μm. Such gratings are difficult to produce as transmission gratings. Therefore reflection gratings are generally used which are produced by carving grooves into a plane glass or quartz surface. Nowadays such gratings are often produced [11] by holographic techniques (see Sect. 12.4).

In order to describe the situation for the reflection, the diffraction and interference we introduce two different normal vectors (Fig. 10.45):

- The grating normal, which is perpendicular to the grating surface
- the groove normal which is perpendicular to the inclined groove surface

When a plane wave incides under the angle α against the grating normal the path difference between the light beams reflected by two adjacent grooves into the direction β against the grating normal is

$$\Delta s = \Delta_1 - \Delta_2 = d(\sin\alpha - \sin\beta), \qquad (10.49a)$$

if the diffraction angle β is on the opposite side of the grating normal as the incident angle α (Fig. 10.45a). If α and β are on the same side of the grating normal we get

$$\Delta s = \Delta_1 + \Delta_2 = d(\sin\alpha + \sin\beta). \qquad (10.49b)$$

In order to describe the two cases by the same formula, the following convention is introduced: The incidence angle α is always positive. The diffraction angle β is positive if it is

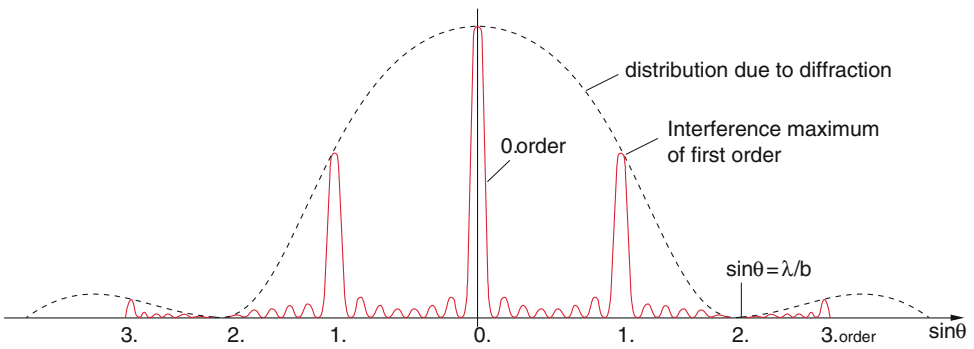


Fig. 10.44 Intensity distribution $I(\theta)$ for a diffraction grating with 8 slits and $d/b = 2$. The second interference order receives no light because the diffraction minimum just falls into this direction

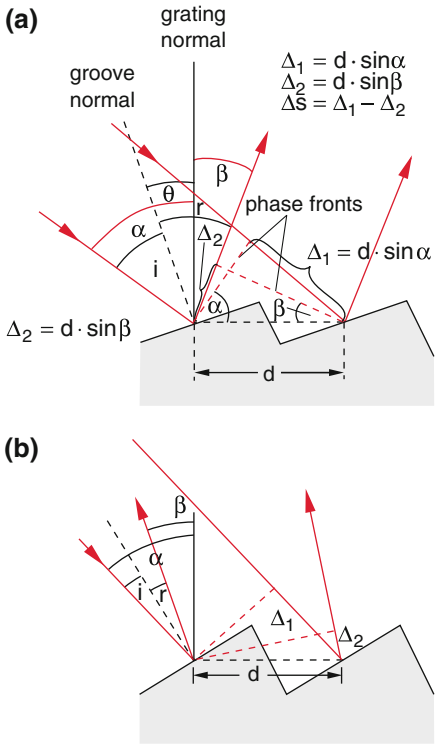


Fig. 10.45 Optical reflection grating. Incident and reflected light are **a**) on different sides of the grating normal. **b**) on the same side

on the same side of the grating normal as α, it is negative if it is on the opposite side. With this definition we can write for both cases:

$$\Delta s = d(\sin\alpha - \sin\beta), \qquad (10.49c)$$

For a given angle α constructive interference of the diffracted intensity is only obtained if the grating equation

$$d(\sin\alpha + \sin\beta) = m \cdot \lambda \qquad (10.50)$$

is fulfilled.

A plane wave incident under the angle i against the *groove* normal is reflected under the angle $r = i$. From Fig. 10.45 we can see that $i = \alpha - \theta$ and $r = \theta - \beta$. (β is negative). For the angle θ between groove normal and grating normal (**Blaze angle**) we therefore get

$$\theta_b = \alpha + \beta/2. \qquad (10.51)$$

The blaze angle gives the inclination of the groove surface against the grating surface. The angle α is generally specified by the construction of the grating spectrograph and the angle β is determined by the groove distance d and the wavelength λ, while the blaze angle θ depends on the

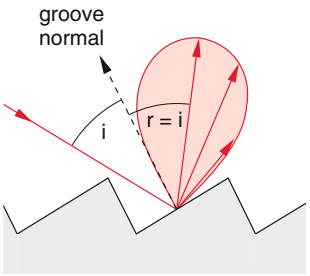


Fig. 10.46 Intensity distribution of light diffracted by a single groove of the grating

inclination of the grooves (Fig. 10.45a). It is, according to (10.50 and 10.51) given by

$$\theta_b = \frac{\alpha}{2} + \frac{1}{2}\arcsin\left[\frac{m \cdot \lambda}{d}\right] \sin\alpha$$

It can be optimized only for a certain wavelength range. It is chosen such, that for the center wavelength λ_m of the wavelength range $\Delta\lambda$ the angle β where the interference maximum of mth order appears, coincides with the reflection angle $r = \theta - \beta$. In this case nearly the whole reflected intensity is concentrated in the mth-order. Because of the diffraction at each groove the reflected light is diffracted into the angular width $\Delta\beta$ around $\beta_m = r - \theta$ (Fig. 10.46). This allows one to detect a wavelength range $\Delta\lambda$ where the intensity $I(\beta)$ varies only slightly.

Example

An optical grating with $d = 1\,\mu m$ is illuminated by parallel light with the wavelength ($\lambda = 0.6\,\mu m$) under the incidence angel $\alpha = 30°$. The first interference order (m = 1) of the reflected light appears, according to (10.50) under the angle β with $\sin\beta = (\lambda - d \cdot \sin\alpha)/d \Rightarrow \sin\beta = 0.1 \Rightarrow \beta \approx +5.74°$. The diffraction angle lies therefore on the other side of the grating normal as the incidence angle α. For $m = -1$ is

$$\sin\beta = -\frac{\lambda}{d} - \sin\alpha = -1.1,$$

this means that the −1. order does not appear. The optimum blaze angle θ_b is then with (10.51) $\theta = 18°$ for $\beta = +6° \Rightarrow r = i = 12°$.

The angular width $\Delta\beta$ of the intensity distribution $I(\beta)$ between the two zero points on both sides of the angle β_1 of the intensity maximum can be obtained from (10.47) for $\theta = \beta$ as $\Delta\beta = \lambda/N \cdot d$.

This complies exactly with the diffraction width of the intensity transmitted through a slit with width $b = N \cdot d$, which equals the width of the whole grating.

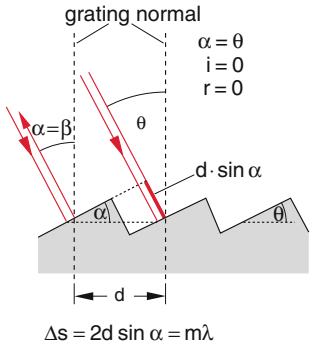


Fig. 10.47 Littrow grating

The intensity distribution of the interference maxima reflected by the grating with N grooves and a total size $N \cdot d$ has the same angular width as the central diffraction maximum of a slit with the same width $b = N \cdot d$.

When the blaze angle θ_b is chosen such that the incident light with wavelength λ falls perpendicular onto the groove surface ($\theta = \alpha$) the mth interference order is reflected back into the incidence direction ($\beta = \alpha$) if the condition

$$\Delta s = 2d \cdot \sin\alpha = m \cdot \lambda.$$

is fulfilled. Such gratings, called **Littrow gratings**, act as wavelength selective mirrors. Even if the incidence angle $\alpha \neq 0$ they reflect the incident intensity back into the incident direction (Fig. 10.47).

10.6 Fraunhofer- and Fresnel-Diffraction

Up to now we have regarded interference- and diffraction phenomena only for parallel incident light beams, which implies the same well defined diffraction angle θ against the direction of the incident light for all partial waves in the beam. This situation is called **Fraunhofer diffraction**. The situation becomes more complicated for divergent or convergent incident light beams where the different parts of the light beam have different incidence angles α within the angular range $\alpha_0 \pm \Delta\alpha$ and therefore experience different diffraction angles θ and different path lengths Δs of the interfering partial waves (**Fresnel Diffraction**).

Remark One can regard Fraunhofer- and Fresnel-Diffraction also as two different approximations of a more general diffraction theory (see Sect. 10.7).

We will illustrate Fresnel diffraction by some examples. At first we will emphasize the importance of Huygens principle by a more detailed discussion of the propagation of a spherical wave.

10.6.1 Fresnel Zones

We regard in Fig. 10.48 a spherical wave which is emitted by the point source L. We will calculate the intensity in an arbitrary point P and determine, how this intensity is altered by obstacles in the way between L and P. On the spherical surface with radius R around L the phase of the wave is constant and the electric field amplitude is

$$E(R) = \frac{E_0}{R} e^{i(\omega t - kR)}. \tag{10.52}$$

Now we regard (following Huygen's principle) every point S on this surface as source of secondary waves. Amplitudes and phases of these secondary waves in the point P depend on the distance SP and the angle θ against the wave vector $\boldsymbol{k}$ of the spherical primary wave in S.

All points S on the spherical surface, which have the same distance $r = SP$ are located on a circle around the line LP with the Radius $\rho = R \cdot \sin\varphi$. With $r(\varphi = 0) = r_0$ we can write the distance $r = LP = R + r_0$. We now construct spheres around P with the radii $r = r_0 + \lambda/2, r_0 + \lambda$; $r_0 + 3/2\lambda$; etc. The intersections of these spheres with the circle around L are circles around the axis LP (dashed curves in Fig. 10.48) which have the distances $r_m = r_0 + m \cdot \lambda/2$ from P. The areas between the circles with distance $r_0 + (m-1)\lambda/2$ and $r_0 + m \cdot \lambda/2$ are called **Fresnel zones**. For each point Q_i inside a Fresnel zone there is a point Q_k in the neighboring zone which has a distance Q_kP that differs by ½λ from the distance Q_iP.

The amplitude E_0 of the light source L has decreased to $E_a = E_0/R$ on the circle with radius R around L. The contribution of the mth Fresnel zone with area dS_m to the field amplitude in P is

$$dE = K \cdot \frac{E_a}{r} e^{i[-k(R+r)+\omega t]}\, dS. \tag{10.53}$$

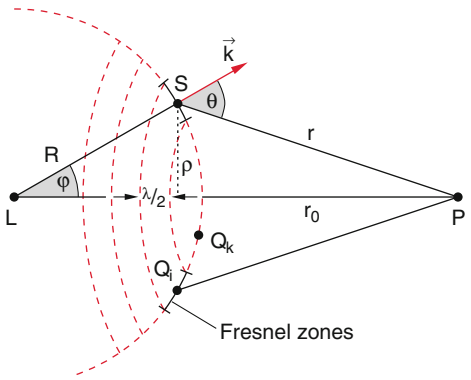


Fig. 10.48 Construction of Fresnel-zones. The Figure is rotationally symmetric about the axis LP

The factor $K(\theta)$ describes the dependence of the field amplitude emitted by dS_m into the direction θ against the surface normal of dS_m. The factor $K(\theta)$ is a slowly varying function of θ (e.g. $K(\theta) = \cos\theta$) and it can be regarded as constant within one Fresnel zone.

$$r^2 = R^2 + (R + r_0)^2 - 2R(R + r_0)\cos\varphi, \tag{10.54}$$

The differentiation with respect to φ gives

$$2r\,dr = 2R(R + r_0)\sin\varphi \cdot d\varphi, \tag{10.55}$$

The area of the Fresnel zone with radius $\rho = R \cdot \sin\varphi$ is

$$dS = 2\pi R \cdot R \cdot \sin\varphi\, d\varphi$$

Inserting $\sin\varphi\, d\varphi$ from (10.55) gives

$$dS = \frac{2\pi R}{R + r_0} r\,dr. \tag{10.56}$$

The contribution of the mth Fresnel zone to the field amplitude in P is then

$$\begin{aligned} E_m &= K_m \cdot E_a \cdot \frac{2\pi R}{R + r_0} \int_{r_{m-1}}^{r_m} e^{-i[k(R+r)-\omega t]}\,dr \\ &= -\left[\frac{\lambda K_m E_a R}{i(R + r_0)} e^{-i[k(R+r)-\omega t]}\right]_{r_{m-1}}^{r_m}. \end{aligned} \tag{10.57}$$

With $k = 2\pi/\lambda$ and $r_m = r_0 + m \cdot \lambda/2$ this becomes

$$E_m = (-1)^{m+1} \frac{2\lambda K_m E_a R}{i(R + r_0)} e^{-i[k(R+r_0)-\omega t]}. \tag{10.58}$$

The contributions of the different Fresnel zones change their sign from one zone to the next. This is obvious, because all points in one zone experience the same phase of the wave emitted by L, but the path length to P changes by $\lambda/2$ from one zone to the next one. The phases of the different partial waves from two adjacent Fresnel zones therefore differ by π.

The total field amplitude $E(P)$ is then

$$\begin{aligned} E(P) &= \sum_{m=1}^{N} E_m \\ &= |E_1| - |E_2| + |E_3| - |E_4| + \cdots \pm |E_N|. \end{aligned} \tag{10.59a}$$

The amounts of E_m vary only slowly with m, because $r \gg \lambda$ and K differs only slightly between neighboring zones (the angle θ barely changes between adjacent zones). We can therefore approximate

$$|E_m| \approx \frac{1}{2}(|E_{m-1}| + |E_{m+1}|). \tag{10.59b}$$

It is therefore reasonable to rearrange the series (10.59a) into

$$\begin{aligned} E(P) = \frac{1}{2}|E_1| &+ \left(\frac{1}{2}|E_1| - |E_2| + \frac{1}{2}|E_3|\right) \\ &+ \left(\frac{1}{2}|E_3| - |E_4| + \frac{1}{2}|E_5|\right) \\ &+ \cdots + \frac{1}{2}|E_N|. \end{aligned} \tag{10.59c}$$

because of (10.59b) all members of this series are negligible except the first and the last term. We therefore obtain

$$E(P) \approx \frac{1}{2}(|E_1| + |E_N|). \tag{10.59d}$$

When we assume, that the factor K is $K = \cos\theta$ the contribution of the last zone with $m = N$, where the line SP is the tangent to the circle around L becomes zero, because $\theta = 90°$ and $\cos 90° = 0$. All zones with $m > N$ cannot emit light into the direction towards P. This gives the final result

$$\begin{aligned} E(P) &\approx \frac{1}{2}E_1 \\ &= \frac{K_1 \lambda E_a R}{i(R + r_0)} e^{-i[k(R+r_0)-\omega t]}. \end{aligned} \tag{10.60}$$

There is also a primary wave propagating from S to P, which contributes to the field amplitude in P. Taking this into account we get

$$E(P) = \frac{E_0}{R + r_0} e^{-i[k(R+r_0)-\omega t]}. \tag{10.61}$$

Of course, (10.60) and (10.61) have to give the same result, because the introduction of a fictive sphere around L and the application of Huygens's principle cannot change the field amplitude in P. The comparison between (10.60) and (10.61) then yields an expression for the factor K which gives with $E_a = E_0/R$ for $m = 1$

$$K_1 = i/\lambda. \tag{10.62}$$

For the mth Fresnel zone is $K_m = i/\lambda \cdot \cos\theta_m$. For the first zone ($m = 1$) is $\theta_1 \approx 0°$ and therefore $\cos\theta_1 \approx 1$.

How large are the Fresnel zones? As can be seen from Fig. 10.48 the radius ρ_m of the mth zone is

$$\rho_m \approx \sqrt{r^2 - r_0^2} = \sqrt{(r_0 + m \cdot \lambda/2)^2 - r_0^2}$$

$$\approx \sqrt{m \cdot r_0 \cdot \lambda} \quad \text{for} \quad r_0 \gg \lambda$$

It therefore depends on the wavelength λ and on the distance r to the observation point P.

Example

$$r_0 = 10\,\text{cm}, \lambda = 0.5\,\mu\text{m} \Rightarrow \rho_1 = 0.22\,\text{mm}.$$

When we place between L and P a screen with an aperture, that equals the diameter $2 \cdot \sqrt{r_0 \lambda}$ of the first Fresnel zone (Fig. 10.49a) the field amplitude transmitted by this aperture is

$$E(P) = E_1 = \frac{2E_0}{R + r_0} e^{-i[k(R+r_0)-\omega t]} \qquad (10.63)$$

This is twice as large as the field amplitude without the screen (the intensity is then 4-times as large). This gives the astonishing result that the introduction of the absorbing screen *increases* the intensity!

The reason is, of course, the prevention of destructive interference between the higher zones and the first zone by the screen. These zones give the contribution $-½E_1$ as can be recognized from the comparison between (10.59a) and (10.59d). The first Fresnel zone acts like a lens, which partially refocuses the divergent light emitted by L.

Instead of selectively transmitting the first Fresnel zone through a circular aperture, one can also selectively block the light from this zone by an absorbing disc (Fig. 10.49b) thus allowing the light from all other zones to reach the detector in P. In the series (10.59a) then the first term is missing. From the rearranged series (10.59c) it can be seen that now the second term is no longer cancelled by the missing first term (because $E_1 = 0$). In this case the total intensity in P is as large as without the absorbing disc.

These surprising facts demonstrate that Huygens's principle (which was postulated by Christiaan Huygens (Fig. 10.50) already in 1690) is very useful to describe the propagation of waves in space, while geometrical optics cannot explain these phenomena.

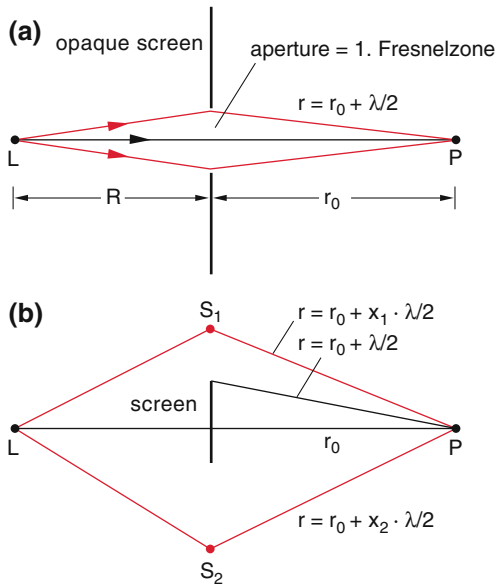


Fig. 10.49 **a**) Through the circular aperture in the opaque screen corresponding to the first Fresnel zone, twice as much intensity is received in P than without the screen. **b**) With the opaque mask which blocks the first Fresnel zone, as much intensity reaches P as without the mask. The points S_i are arbitrary points in the plane $z = 0$

Fig. 10.50 Christiaan Huygens (1629–1695). (With kind permission of "Deutsches Museum München")

When the distance R in Fig. 10.49 between light source L and aperture becomes very large compared with the diameter of the aperture, the incident wave can be regarded as plane wave and the virtual sphere in Fig. 10.48 with the Fresnel zones converges against a plane surface (Fig. 10.51). The radius ρ_m of the mth Fresnel zone still depends on the distance r_0 of the observation point P.

Fresnel diffraction is always observed when the aperture which contributes to the illumination in P contains many Fresnel zones, i.e. when its diameter $D \gg \sqrt{r_0 \cdot \lambda}$. This implies that many Fresnel zones contribute to the field amplitude in P. If r_0 is so large, that only the first Fresnel zone contributes one obtains Fraunhofer diffraction patterns [11].

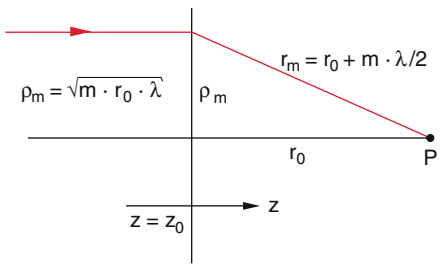


Fig. 10.51 Radius of the mth-Fresnel zone in the plane $z = z_0$ of a plane wave propagating into the z-direction for an observer in P at $z = z_0 + r_0$

10.6.2 Fresnel's Zone Plate

The result of the foregoing section can be utilized to concentrate more light onto the observation point P, as is possible with the simple circular aperture in Fig. 10.49a. For this purpose one uses instead of the screen a glass plate where opaque circular rings are vapor deposited which block all Fresnel zones with odd zone number m (Fig. 10.52). This arrangement transmits the light from all zones with even m. in the series (10.59a). Therefore only terms with the same sign appear which are all in phase. This eliminates destructive interference.

Such an arrangement is called **Fresnel's zone plate**. The diameter and the width of the transmitting rings depend on the distance LP and on the distance r_0 between zone plate and observation point P. As can be inferred from Fig. 10.51 the radius of the mth zone for $r_0 \gg m \cdot \lambda$ is obtained from the relation $\rho_m^2 = (r_0 + m \cdot \lambda/2)^2 - r_0^2 \Rightarrow \rho_m^2 = r_0 m\lambda + m^2\lambda^2/4$. With $r_0 \gg m \cdot \lambda$ this can be simplified to

$$\rho_m = \sqrt{m r_0 \cdot \lambda}. \tag{10.64}$$

The width of the mth zone

$$\begin{aligned} \Delta\rho_m &= \rho_{m+1} - \rho_m \\ &= \sqrt{r_0\lambda}\left(\sqrt{m+1} - \sqrt{m}\right) \end{aligned} \tag{10.65}$$

decreases with increasing m. However, the area of the zones

$$A_m = \pi(\rho_{m+1}^2 - \rho_m^2) = \pi r_0 \lambda \tag{10.66}$$

is the same for all zones.

Such a zone plate acts as a lens that collects light which is incident onto the plate within a certain angular range (**Fresnel lens**).

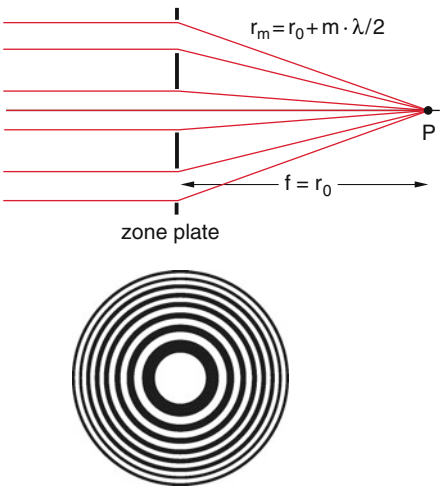


Fig. 10.52 Fresnel's zone plate

If r_0 is the distance of the point P from the center of the zone plate, then all points of the mth zone have the distance $r = r_0 + m \cdot \lambda/2$. from P. When the zone plate is illuminated with a parallel light beam from the left side in Fig. 10.52, the secondary waves in all "open" zones are excited in phase. Since the path ways from the open zones differ by λ between adjacent open zones all secondary waves arrive in P with equal phase, the point P is therefore the focal point of the incident wave and the focal length $f = r_0$ is obtained from (10.64) as

$$f = \frac{\rho_m^2}{m \cdot \lambda} = \frac{\rho_1^2}{\lambda}. \tag{10.67}$$

The focal length of a Fresnel zone plate is given by the radius ρ_1 of the first zone and the wavelength λ. A zone plate therefore has a wavelength-dependent focal length.

Such zone plates that represent lenses have gained increasing importance for imaging wavelength ranges where no transparent material for classical lenses is available. This applies in particular to the X-ray region where zone plates are the only possible lenses. Glass or quartz lenses not only absorb X-rays but have also a refractive index $n \approx 1$ which implies that their focusing properties are nearly zero.

The experimental realization of Fresnel lenses for X-rays uses a thin foil, transparent for X-rays. The opaque zones are realized by vapor deposition of heavy metals [12]. Special Fresnel lenses with zones that are transparent for atoms are also used for focusing mono-energetic atomic beams (see Vol. 3).

10.7 General Treatment of Diffraction

We will now discuss a general way how to describe and calculate diffraction by apertures or obstacles of arbitrary form. Although such calculations are often only possible with numerical methods, the simplified version of the Fresnel-Kirchhoff diffraction theory, represented here, can give a better insight into the basic ideas of the Fresnel diffraction.

10.7.1 The Diffraction Integral

We regard in Fig. 10.53 an arbitrary hole with the area σ in a screen, which is placed in the x-y-plane at $z = 0$ and which is illuminated by a light wave. We will calculate the intensity distribution in the x'-y'-plane at $z = z_0$ (observation plane). The electric field amplitude can be described by

$$E_S(x, y) = E_0(x, y) \cdot e^{i\varphi(x,y)} \tag{10.68}$$

A point-like light source placed at $L = (0, 0, -g)$ emits its radiation with the amplitude A uniformly into all directions (Fig. 10.53b). At the position of the screen we get the amplitude

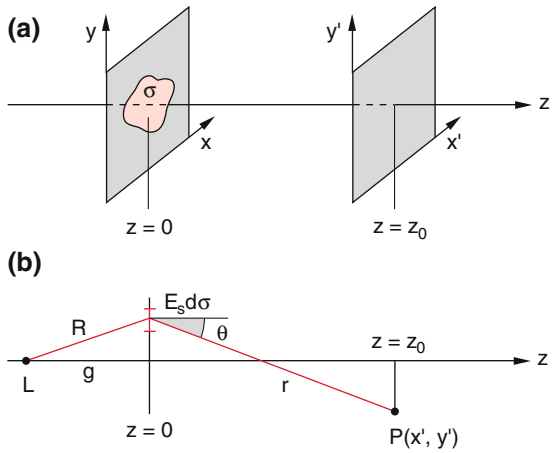


Fig. 10.53 Illustration of the derivation of the Fresnel-Kirchhoff-diffraction integral

$$E_0(x,y) = \frac{A}{R} = \frac{A}{\sqrt{g^2 + x^2 + y^2}} \quad \text{and} \qquad \varphi = (\omega t - kR). \tag{10.68a}$$

The infinitesimal area $d\sigma$ of the hole emits according to Huygens's principle secondary waves which contribute to the field amplitude in the point $P(x', y')$ the amount

$$\mathrm{d}E_P = C \cdot \frac{E_S \cdot \mathrm{d}\sigma}{r} \mathrm{e}^{-ikr} \tag{10.69}$$

As has been discussed in Sect. 10.6.1 the proportional factor C can be written as $C = i \cdot \cos\theta/\lambda$.

The total radiation of the illuminated hole at $z = 0$ generates at the point P the field amplitude

$$E_P = \int\int C \cdot E_S \cdot \frac{\mathrm{e}^{-ikr}}{r} \mathrm{d}x\,\mathrm{d}y, \tag{10.70}$$

where the two-dimensional integral extends over all area elements $\mathrm{d}\sigma = \mathrm{d}x \cdot \mathrm{d}y$ of the hole in the screen. The integral (10.70) is called **Fresnel-Kirchhoff diffraction integral**.

If the distance r between the points $S(x,y)$ and the observation point $P(x', y')$ is large compared with the distances x, y of the hole elements at $x/z_0 \ll 1$ and $y/z_0 \ll 1$) we can replace the distance r in the denominator in (10.70) by $r \approx z_0$. The phase in the exponent depends, however, sensitively on r and therefore we cannot replace here r by z_0 but have to use a better approximation. In the Taylor expansion

$$r = \sqrt{z_0^2 + (x - x')^2 + (y - y')^2} \approx z_0 \left(1 + \frac{(x - x')^2}{2z_0^2} + \frac{(y - y')^2}{2z_0^2} + \cdots\right) \tag{10.71}$$

we keep all terms up to the quadratic one and neglect only the higher order terms. With $\cos\theta = z_0/r \approx 1 \Rightarrow C = (i/\lambda)$ the diffraction integral becomes

$$E(x', y', z_0) = \mathrm{i}\frac{\mathrm{e}^{-ikz_0}}{\lambda z_0} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E_S(x,y) \cdot \exp\left[\frac{-ik}{2z_0}\left((x - x')^2 + (y - y')^2\right)\right] \mathrm{d}x\,\mathrm{d}y. \tag{10.72}$$

This formula allows one to calculate the distribution of the electric field amplitude $E(x', y', z_0)$, if the field distribution $E(x,y)$ in the plane $z = 0$ is known.

The approximation, used in the derivation of (10.72) is called **Fresnel-approximation**. If the diameter of the hole is small compared to z_0, a further approximation can be used. With

$$z_0 \gg \frac{1}{\lambda}\left(x^2 + y^2\right),$$

the quadratic terms x^2 and y^2 in (10.71) can be also neglected and we get

$$r \approx z_0 \left(1 - \frac{xx'}{z_0^2} - \frac{yy'}{z_0^2} + \frac{x'^2 + y'^2}{2z_0^2}\right).$$

Since the integration extends over x and y we can extract the terms with x' and y' out of the integral and we obtain instead of (10.72) the expression

$$E(x', y', z_0) = A(x', y', z_0) \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E_S(x,y) \cdot \exp\left[\frac{+ik}{z_0}(x'x + y'y)\right] \mathrm{d}x\mathrm{d}y \tag{10.73}$$

with

$$A(x', y', z_0) = \frac{\mathrm{i}\mathrm{e}^{-ikz_0}}{\lambda z_0} \cdot \mathrm{e}^{(-i\pi)/(\lambda z) \cdot (x'^2 + y'^2)}.$$

This approximation is the **Fraunhofer-diffraction**, where the diffraction structures are observed in the far-field region. The general case, where the linear approximation is no longer valid is called **Fresnel diffraction**.

We will now illustrate the two cases by some examples.

10.7.2 Fresnel- and Fraunhofer Diffraction by a Slit

A narrow slit in y-direction with the width $\Delta x = b \gg \lambda$ should be illuminated by a parallel light beam (Fig. 10.54). We will determine the intensity distribution $I(x', z_0)$ of the diffracted light in the plane $y = 0$ for different distances z_0 from the plane $z = 0$ of the slit. The diffraction integral (10.72) reduces to a one-dimensional integral

$$E(P) = C \cdot E_S \int_{-b/2}^{+b/2} \frac{1}{r} e^{-ikr} dx, \qquad (10.74)$$

where

$$r = [(x - x')^2 + z_0^2]^{1/2} = z_0 \sqrt{1 + \left(\frac{x - x'}{z_0}\right)^2}$$

is the distance from the observation point P to a point $(x, 0, 0)$ of the slit. The field amplitude E_S is constant over the whole slit and can be therefore extracted before the integral. We distinguish between three different observation zones.

- the near-field zone where z_0 is of the same order of magnitude as the slit width $b \gg \lambda$. Here the radius $r_1 = \sqrt{(z_0 \cdot \lambda)}$ of the first Fresnel zone is small compared with the slit width b and many Fresnel zones contribute to the field amplitude in P. This means that the phase of the total wave in P strongly varies with x'. The numerical integration of (10.74) gives the intensity distribution $I(x) \propto |E(x)|^2$ shown in the left part of Fig. 10.54.
- the medium distance zone, where only a few Fresnel zones contribute to $I(x')$ is shown in the middle part of Fig. 10.54.
- the far field zone ($z_0 \gg b$) where the radius $r_1 = \sqrt{z_0 \lambda}$ of the first Fresnel zone is larger than b. This is the region of the Fraunhofer diffraction (right picture in Fig. 10.54, where the approximation (10.73) is valid. All terms that do not depend on x and also the essentially constant denominator r can be extracted from the integral. This gives the Fraunhofer diffraction formula for the diffracted intensity distribution (10.43) (see Problem 10.5).

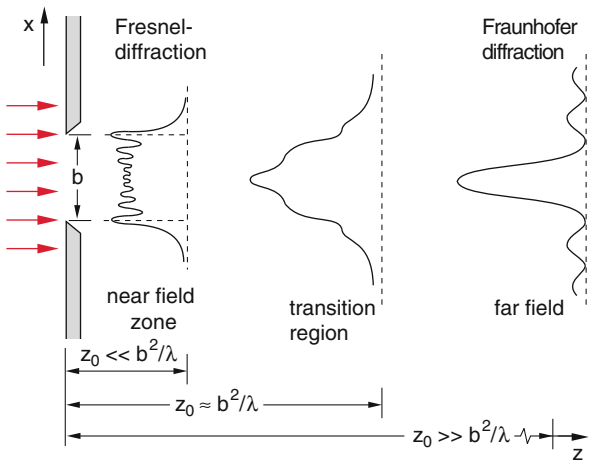


Fig. 10.54 Fresnel- and Fraunhofer diffraction behind a slit. Illustrated are the intensity distributions for different distances behind the slit in the "near" and the "far-zone"

The discussion above illustrates that the intensity distribution $I(\theta)$ of Fraunhofer-diffraction, which commonly represents the diffraction by a slit, is an approximation which is only valid for sufficiently large distances ($r \gg b$) of the observation point P behind the diffracting slit, but dos not describe the observed phenomena in the near field zone.

The infinitely far away observation point of the far field can be transferred by a lens behind the diffracting aperture into the focal plane of this lens. The focal length of the lens has to be large compared to the diameter of the aperture.

10.7.3 Fresnel Diffraction at an Edge

When a parallel light beam incides onto an opaque screen in the x-y-plane at $z = 0$ which covers the plane $x < 0$ with an edge along the y-axis ($x = 0$) one observes the diffraction pattern shown in Fig. 10.55. There is also some light penetrating into the half-space $x' < 0$ where without diffraction no light should be present, while in the half-space $x' > 0$ an oscillating intensity $I(x')$ can be seen.

The diffraction integral (10.47) in the observation point P becomes

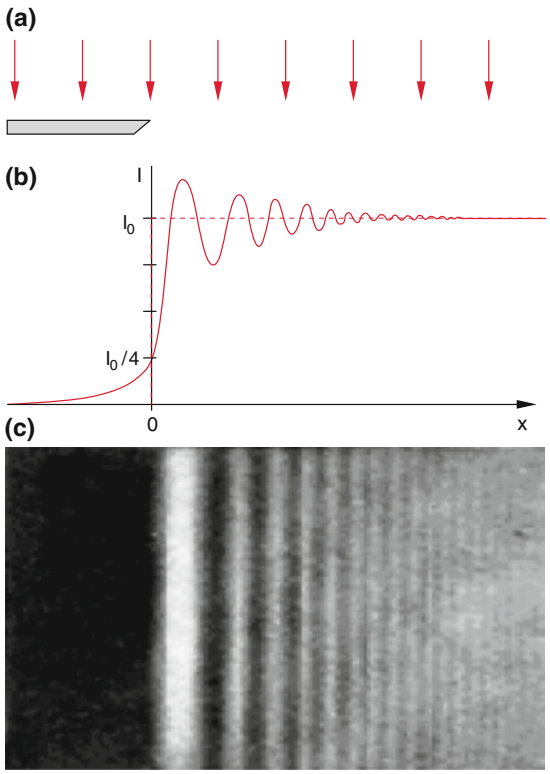


Fig. 10.55 Intensity distribution behind a diffracting edge. **a**) Schematic drawing, **b**) calculated from the diffraction integral. The dashed curve gives the intensity without diffraction, **c**) observed intensity distribution (From D. Meschede: Gehrtsen, Physik, 21 ed. Springer Berlin, Heidelberg)

$$E(P) = C \cdot E_S \int_0^\infty \frac{e^{-ik\sqrt{(x-x')^2 + z_0^2}}}{\sqrt{(x-x')^2 + z_0^2}} dx. \qquad (10.75)$$

For $x' \ll z_0$ the integral can be solved by a series expansion [13] and gives the diffraction intensity pattern $I(x')$, shown in Fig. 10.55.

10.7.4 Fresnel Diffraction at a Circular Aperture

When a circular aperture with radius a in an opaque screen is illuminated by a parallel light beam one observes in the plane $z = z_0$ behind the aperture a diffraction structure that has rotational symmetry around the z-axis (Fig. 10.56). The intensity distribution $I(\rho)$ with $\rho^2 = x'^2 + y'^2$ depends on the diameter $2a$ of the aperture and the distance z_0 between observation point P and the aperture. The intensity $I(\rho = 0)$ in the central point $P_0(\rho = 0)$ becomes maximum for $z_0 = a^2/\lambda$, because then the area of the first Fresnel zone with radius $r_1 = \sqrt{z_0 \cdot \lambda} = a$ equals the area of the circular aperture (see Sect. 10.6.1). For a smaller distance $z_0 = a^2/2\lambda$ (or a larger aperture radius) the aperture area covers the first two Fresnel zones. Their contributions to the field amplitude in P_0 interfere destructively, which decreases the intensity in P_0 nearly to zero. One observes a dark central point of the circular diffraction pattern.

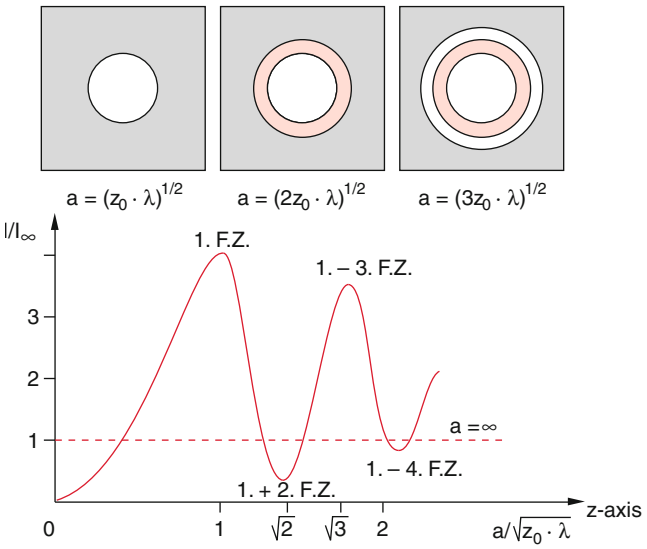


Fig. 10.56 Intensity distribution of light in a point $P(z_0)$ due to diffraction by a circular aperture, as a function of the aperture radius a. The upper part illustrates the Fresnel zones for aperture with $a = \sqrt{nz_0 \cdot \lambda}$ for $n = 1, 2, 3$, corresponding to the maxima in the lower part. The light passing through the 2. Fresnel zone has a path difference of $\lambda/2$ and interferes destructively. The dashed lie gives the intensity without diffraction ($a = \infty$)

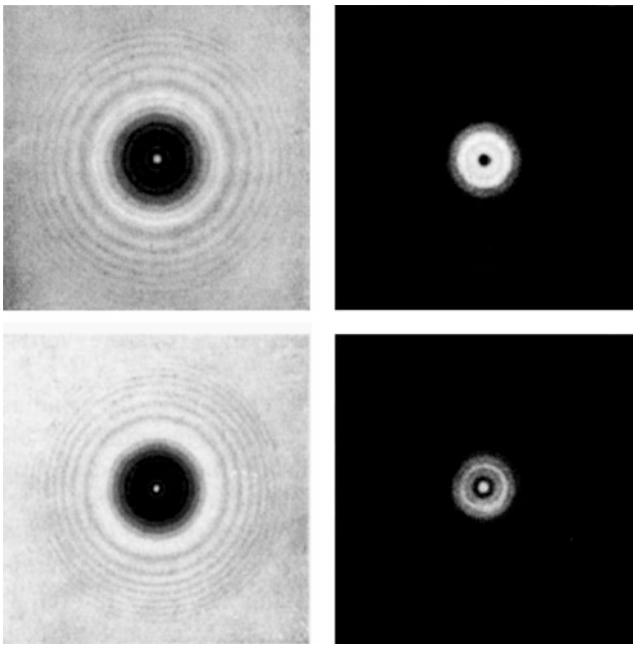


Fig. 10.57 Comparison of the diffraction pattern by a circular aperture (right) and a opaque disc with equal diameter (left). The photos in the upper and lower part have been taken at different distances from the aperture resp. disc (From Weizel, Lehrbuch der theoretischen Physik, Springer Berlin, Heidelberg 1949)

The intensity $I(P_0)$ varies periodically along the z-axis when the distance from the aperture is altered.

A similar diffraction pattern is observed when the screen with the circular aperture is replaced by a circular opaque disc with radius a (Fig. 10.57). For this case one also observes maximum intensity on the z-axis for $z_0 = a^2/\lambda$ and minimum intensity for $z_0 = a^2/2\lambda$.

10.7.5 Babinet's Theorem

From Eq. (1.72) we see, that the electric field strength $\boldsymbol{E}_P$ in the observation point P is determined by the surface integral of the electric field amplitude over the area σ of the aperture. The calculation of the diffraction phenomena caused by apertures or obstacles with a more complicated form is facilitated by a theorem, first postulated by *J. Babinet (1794–1872)*. It is based on the following statements:

If the area σ of the aperture is divided into two subareas σ_1 and σ_2 the field amplitude measured in P is

$$E_P(\sigma) = E_P(\sigma_1) + E_P(\sigma_2) \ ,$$

where $E_P(\sigma_i)$ is the field amplitude which would be measured if the aperture only includes the sub-area σ_i. The more general statement is:

If the area σ is divided into N subareas the total field amplitude in P is

$$E_P(\sigma) = \sum_{i=1}^{N} E_P(\sigma_i). \qquad (10.76)$$

Examples

1. An annulus aperture with inner radius ρ_1 and outer radius ρ_2 generates in P a field amplitude $E_P = E_P^{(1)} - E_P^{(2)}$ where $E_P^{(i)}$ is the field amplitude that is generated by a circular disc with radius ρ_i. Of course one has to consider the different phases of $E_P^{(1)}$ and $E_P^{(2)}$ in P.
2. A rectangular aperture is divided into two sub-areas as shown in Fig. 10.58a. The diffraction pattern of the complex form σ_1 can be obtained as the difference

$$E_P(\sigma_1) = E_P(\sigma) - E_P(\sigma_2)$$

between the patterns generated by the simpler areas σ and σ_2 which are much easier to calculate.

Two apertures σ_1 and σ_2 are called complementary to each other when σ_1 is transparent at such parts of the aperture where σ_2 is opaque. Further examples of complimentary apertures are a circular hole in an opaque screen and an opaque circular disc or a slit with width b and a straight wire with a diameter $d = b$ (Fig. 10.58b, c). For the cases (b) and (c) the sum $\sigma_1 + \sigma_2$ gives the total unlimited area which shows no diffraction effects because it has no edges. For the total field amplitude we therefore get

$$E_P(\sigma_1) = -E_P(\sigma_2)\ . \qquad (10.77)$$

For the intensity distribution $I(P) = |(E_P)|^2$ one gets the astonishing result that the diffraction intensity pattern of an opaque circular disc and of a transparent circular aperture are equal, if one subtracts the intensity from S that reaches P through the aperture on a geometrical path (i.e. without diffraction).

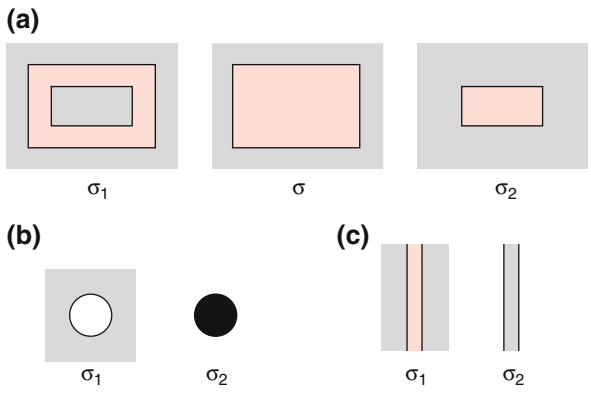


Fig. 10.58 Complementary diffraction areas: **a**) rectangular aperture, **b**) circular aperture and opaque circular disc of the same size, **c**) slit and wire with equal thickness

10.8 Fourier Representation of Diffraction

Using the Fourier-Theorem the diffraction at arbitrarily formed apertures can be described quite generally in a mathematically elegant form. This has considerably advanced modern optics. We will therefore shortly discuss the basic principles of Fourier-optics.

10.8.1 Fourier-Transformation

For an arbitrary real or complex function $f(x)$ which is square-integrable the integral

$$\int_{-x_0}^{+x_0} |f(x)|^2 \mathrm{d}x$$

must remain finite for $x_0 \to \infty$. The Fourier-transform of $f(x)$ is defined as the function

$$F(u) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x) \cdot \mathrm{e}^{-\mathrm{i}ux} \mathrm{d}x. \qquad (10.78)$$

In order to calculate $f(x)$ from $F(u)$ we multiply both sides of (10.78) with $\mathrm{e}^{\mathrm{i}2\pi ux'}$ and integrate both sides over the variable u. This gives, when we subsequently rename x' by x

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(u) \mathrm{e}^{+\mathrm{i}ux} \mathrm{d}u. \qquad (10.79)$$

The two functions $f(x)$ and $F(u)$ are called a Fourier-pair and the variables x and u are Fourier-conjugated variables. The dimensional units of x and u must be reciprocal, because the product $u \cdot x$ in the exponent has to be dimensionless.

Example

The frequency spectrum $F(\omega)$ of the exponentially decreasing light amplitude

$$E(t) = A_0 \cdot \mathrm{e}^{-\gamma t} \cos \omega_0 t \qquad (10.80)$$

can be obtained from (10.78) with $u = \omega$; $x = t$ and $E(t) = f(x)$. This gives for the initial condition $A_0(t) = 0$ for $t < 0$.

$$F(\omega) = \frac{A_0}{\sqrt{\pi/2}} \int_0^{+\infty} \mathrm{e}^{-\gamma t} \cos \omega_0 t \cdot \mathrm{e}^{-\mathrm{i}\omega t} \mathrm{d}t. \qquad (10.81)$$

The integral can be readily solved and gives for $\omega \gg (\omega_0 - \omega)$

$$F(\omega) = \frac{\gamma A_0}{\sqrt{2\pi}} \frac{1}{(\omega_0 - \omega)^2 + \gamma^2}. \qquad (10.82)$$

$F(\omega)$ is the amplitude of the light wave at the frequency ω. The frequency spectrum of the intensity $I \propto A \cdot A^*$ is the Lorentz-profile

$$I(\omega) = \frac{C}{\left[(\omega_0 - \omega)^2 + \gamma^2\right]^2}, \qquad (10.83)$$

where the constant C is chosen such, that the integral $\int I(\omega) d\omega$ is equal to the total intensity I_0 (Fig. 10.59).

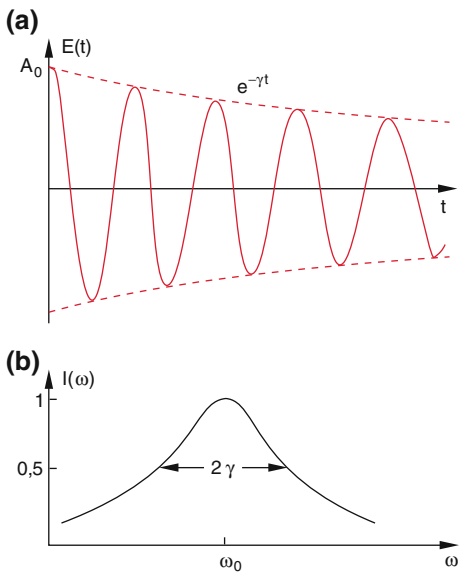


Fig. 10.59 **a**) Temporally decaying amplitude $E(t)$ of a light wave. **b** Fourier-transform $I(\omega)$ of $EE^*(t)$

For the representation of diffraction theory one needs the two-dimensional Fourier-transformation

$$F(u,v) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(x,y) \cdot \mathrm{e}^{-\mathrm{i}2\pi(u \cdot x + vy)} \mathrm{d}x\, \mathrm{d}y, \qquad (10.84a)$$

$$f(x,y) = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} F(u,v) \cdot \mathrm{e}^{\mathrm{i}2\pi(u \cdot x + v \cdot y)} \mathrm{d}u\, \mathrm{d}v. \qquad (10.84b)$$

If the function $f(x, y)$ can be split into two factors ($f(x, y) = f_1(x) \cdot f_2(y)$ then the Fourier-transform

$$F(u,v) = F_1(u) \cdot F_2(v), \qquad (10.85)$$

can be also split into two one-dimensional functions $F_1(u)$ and $F_2(v)$, where $F_1(u)$ is the Fourier-transform of $f_1(x)$ and $F_2(v)$ that of $f_2(y)$.

10.8.2 Application to Diffraction Problems

We will now treat the general case that a light wave with the field amplitude $E_i(x, y)$ falls onto an area σ in the plane $z = 0$ with the transmission $\tau(x, y)$. For an aperture, for instance, is $\tau(x, y) = 1$ inside the aperture opening and $\tau = 0$ outside (Fig. 10.60).

Directly behind the area σ is

$$E(x,y) = \tau(x,y) \cdot E_i(x,y). \qquad (10.86)$$

The spatial distribution of the field amplitude in the observation plane $z = z_0$ can be calculated when using the diffraction integral (10.73). Inserting (10.84a, 10.84b) into (10.73) and comparing the result with (10.84a, 10.84b) where one has to replace $u = x'/(\lambda z_0)$ and $v = y'/(\lambda z_0)$ one recognizes that

$$f(x,y) = E(x,y) = \tau(x,y) \cdot E_i(x,y) \qquad (10.87)$$

describes the amplitude distribution directly behind the diffraction plane $z = 0$.

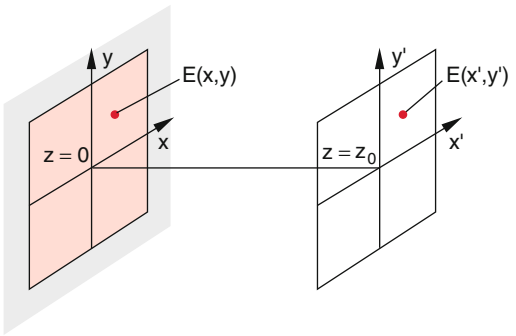


Fig. 10.60 Fourier-representation of Fraunhofer diffraction

The field amplitude $E(x', y')$ in the observation plane $z = z_0$ is then with (10.73)

$$E(x', y') = A(x', y', z_0) \cdot \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} E_e(x, y) \cdot \tau(x, y) \cdot e^{-i2\pi(x'x + y'y)/(\lambda z_0)} dx\, dy. \tag{10.88}$$

The comparison with (10.84a, 10.84b) then yields

$$E(x', y', z_0) = F(u, v) \cdot A(x', y', z_0). \tag{10.89a}$$

We therefore obtain the important result:

The amplitude distribution of the Fraunhofer-diffraction pattern in the observation plane $z = z_0$ is proportional to the Fourier-transform $F(x', y')$ of the function $f(x, y) = \tau(x, y) \cdot E_i(x, y)$ where $\tau(x, y)$ is the transmission function of the diffracting area and $E_i(x, y)$ the field distribution of the incident wave.

The intensity distribution in the observation plane is then

$$I(x', y') \propto |E(x', y')|^2 = |A(x', y')|^2 \cdot |F(x', y')|^2 \tag{10.89b}$$

because $|A(x', y')|^2 = 1$.

We will apply this result onto the diffraction at a rectangular opening. Further examples follow in Sect. 12.5.

Rectangular Aperture

We regard in Fig. 10.61 a rectangular aperture with width a and height b in an opaque screen. The transmission function $\tau(x, y)$ is then

$$\tau(x, y) = \begin{cases} 1 & \text{for } -a/2 < x < +a/2, \\ & -b/2 < y < b/2, \\ 0 & \text{otherwise.} \end{cases}$$

A plane extended wave $E_i(x, y) = E_0 = \text{const.}$ incides onto the aperture. We divide the aperture into small stripes with width dx. The contribution $dE(x', y')$ of such a stripe to the field amplitude is according to (10.88)

$$dE(x', y') = E_0 e^{-2\pi i x' x/(\lambda z_0)} dx \cdot \int_{-b/2}^{+b/2} e^{-2\pi i y' y/(\lambda z_0)} dy. \tag{10.90a}$$

Integration over all stripes yields the electric field distribution in the observation plane

$$E(x', y') = E_0 \cdot \int_{-a/2}^{+a/2} e^{-2\pi i x' x/(\lambda z_0)} dx. \int_{-b/2}^{+b/2} e^{-2\pi i y' y/(\lambda z_0)} dy. \tag{10.90b}$$

The integration gives

$$E(x', y') = E_0 \cdot \frac{\lambda^2 z_0^2}{\pi^2 x' y'} \cdot \sin\frac{\pi x' a}{\lambda z_0} \cdot \sin\frac{\pi y' b}{\lambda z_0}. \tag{10.91}$$

The intensity in the observation plane $I(x', y') \propto |A|^2 \cdot |E|^2$ is then

$$I(x', y') = I_0 \cdot \frac{\sin^2(\pi x' a/\lambda z_0)}{(\pi x' a/\lambda z_0)^2} \cdot \frac{\sin^2(\pi y' b/\lambda z_0)}{(\pi y' b/\lambda z_0)^2}. \tag{10.92}$$

The comparison with (10.45) shows with $\sin\theta = x'/z_0$ resp. y'/z_0 the same result, which had been derived in a completely different way. A rectangular aperture $a \cdot b$ has therefore a diffraction pattern which is equal to that of two mutual perpendicular infinitely long slits in x- and y-direction with widths a and b.

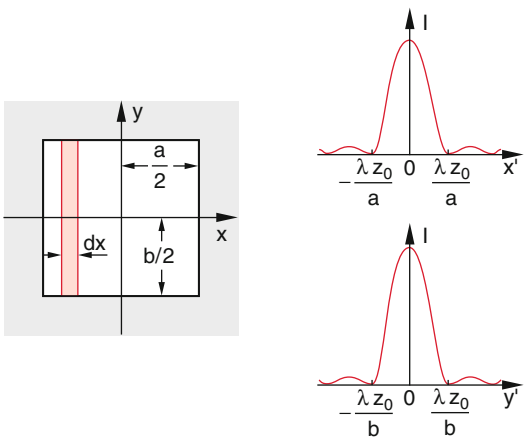


Fig. 10.61 Diffraction at a rectangular aperture

10.9 Light Scattering

In Sects. 8.1 and 8.2 dispersion and absorption were explained by the interaction of the electro-magnetic wave with atomic oscillators which are induced to forced oscillations in the direction of the $\boldsymbol{E}$-vector of the wave. Each dipole radiates the average power

$$\overline{P}_S = \frac{e^2 x_0^2 \omega^4}{32\pi^2 \varepsilon_0 c^3} \sin^2\vartheta \tag{10.93}$$

into the solid angle $d\Omega = 1$ sterad around the direction with the angle ϑ against the dipole axis. A plane wave polarized

in the x-direction which travels into the z-direction aligns the dipoles and induces oscillations in the x-direction. These oscillating dipoles emit radiation according to (10.93) also into directions that deviates by the angle $\alpha = \pi/2 - \vartheta$ from the z-direction. This phenomenon, that light, induced by the incident wave in z-direction, is emitted into all directions, is called **light scattering** [14, 15].

The following questions arise:

- Under which conditions can light scattering be observed?
- What is the frequency dependence of light scattering?
- Which quantities determine light scattering?
- Why does a light beam propagating through a homogeneous medium not suffer any scattering, although the atomic dipoles emit their radiation into all directions?

10.9.1 Coherent and Incoherent Scattering

In Sect. 10.5.1 it was shown that for N oscillators which oscillate in phase, the total intensity is emitted only into those directions where the contributions of the different dipoles show constructive interference i.e. their radiation superimposes in phase. We call the scattering by phase-coupled oscillators **coherent scattering**. One example is the propagation of a plane wave through a homogeneous crystal with regularly arranged atoms (Fig. 10.62) where the wave traverses straight on without any scattering, if the distance d between the atoms is small compared to the wavelength λ. ($d/\lambda \ll 1$) but the total width of the crystal perpendicular to the propagation direction is large compared to the wavelength ($D = dN^{1/3} \gg \lambda$) (so that diffraction effects can be neglected).

The situation changes dramatically if the atoms are irregularly placed (e.g. in powder) or if they perform irregular thermal motions and therefore the distances between the atoms change statistically in time (for example in a liquid or a gas). In such cases there is no longer a fixed phase relation

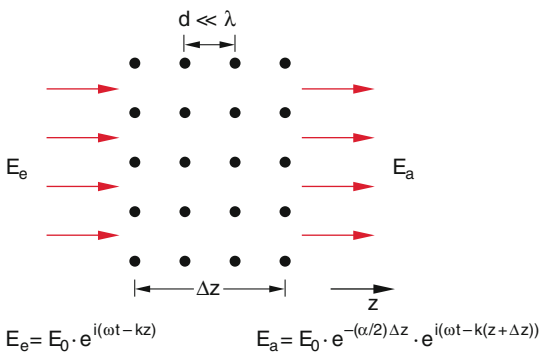


Fig. 10.62 An electro-magnetic wave propagating through an ideal crystal with atom distance $d \ll \lambda$ suffers a phase delay but no scattering

between the emission of the different oscillators as for the coherent scattering and there is no well-defined coherent superposition of the emission from the different atoms. We call this situation **incoherent scattering**.

We will illustrate the difference by the simple example of the superposition of the radiation from two oscillators located at the positions $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ with distance d and oscillation amplitudes (Fig. 10.63).

$$x_1(t) = A_1 \cdot \cos \omega t; \quad x_2(t) = A_2 \cdot \cos(\omega t + \varphi)$$

The total intensity in the direction α is then

$$I = c\varepsilon_0 [A_1 \cdot \cos \omega t + A_2 \cdot \cos(\omega t + \psi)]^2, \tag{10.94}$$

where the phase shift is

$$\psi = \varphi + 2\pi/\lambda d \cdot \sin \alpha$$

The total phase shift ψ is the sum of the temporal phase shift φ between the two oscillators and the spatial phase difference $(2\pi/\lambda)d \cdot \sin \alpha$. Calculating the square of the bracket in (10.94) yields with the relation $2 \cos \alpha \cdot \cos \beta = \cos(\alpha + \beta) + \cos(\alpha - \beta)$

$$\begin{aligned} I = c\varepsilon_0 [&A_1^2 \cos^2(\omega t) + A_2^2 \cos^2(\omega t + \psi) \\ &+ A_1 A_2 (\cos(2\omega t + \psi) + \cos \psi)]. \end{aligned} \tag{10.95a}$$

All detectors available today cannot follow the rapid oscillation of light but measure the time average of the intensity, which is because of $\overline{\cos^2 \omega t} = 1/2$ and $\overline{\cos \omega t} = 0$

$$\bar{I} = \frac{1}{2} c\varepsilon_0 [A_1^2 + A_2^2 + 2A_1 A_2 \overline{\cos \psi}] \tag{10.95b}$$

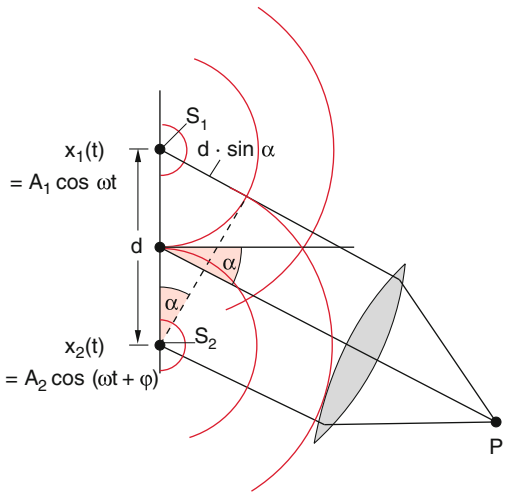


Fig. 10.63 Superposition of the scattering amplitudes in the direction α, caused by two scattering particles S_1 and S_2 with the mutual distance d. **Note** that the plane through S_1 and S_2 is no longer a phase plane for light scattered in the direction $\alpha \neq 0$

here we have anticipated possible temporal fluctuations of the of the phase ψ are slow compared to the oscillation period $T = 2\pi/\omega$. For a temporal constant phase ψ (phase coupled oscillators) is $\overline{\cos\psi} = \cos\psi$. Then the time-averaged intensity depends on the phase ψ and can vary between the maximum intensity

$$I_{\max} = 1/2c\varepsilon_0(A_1 + A_2)^2$$

for

$$\psi = m \cdot 2\pi, \text{ with } m = 0, 1, 2, \ldots \tag{10.96a}$$

and the minimum intensity

$$I_{\min} = \frac{1}{2}c\varepsilon_0(A_1 - A_2)^2$$

for

$$\psi = (2m+1)\pi \tag{10.96b}$$

There are interference phenomena (see Sect. 10.4 and Vol. 1, Sect. 11.10), which result in spatial structures of the intensity distribution (coherent superposition) For coherent scattering one can observe a spatially varying intensity which shows for $d > \lambda$ maxima for certain angles α against the direction of the incident parallel light beam.

For the case of incoherent scattering by particles with a mean distance $d > \lambda$ the phase ψ varies statistically between $-\pi$ and $+\pi$ and the time average of $\overline{\cos\psi} = 0$. Therefore the time average of the total intensity becomes

$$\overline{I} = \frac{1}{2}c\varepsilon_0(A_1^2 + A_2^2). \tag{10.97}$$

If, for example, the distances between the scattering particles are randomly distributed in space, also the phases of these oscillators excited by a plane wave are randomly distributed, which causes the time average of cos ψ to vanish ($\overline{\cos\psi} = 0$).

We therefore obtain the important result:

For the *coherent scattering* the total intensity is the square of the sum of the different amplitudes (taking into account the relative phases). For the *incoherent scattering* the intensities of the different contributions are added. The relative phases are not important because the average to zero.

The time average of the total intensity, incoherently scattered by N particles into the solid angle $d\Omega = 1$ sr around the angle ϑ against the electric vector $\boldsymbol{E}$ of the incident wave (Fig. 10.64) is then according to (10.63)

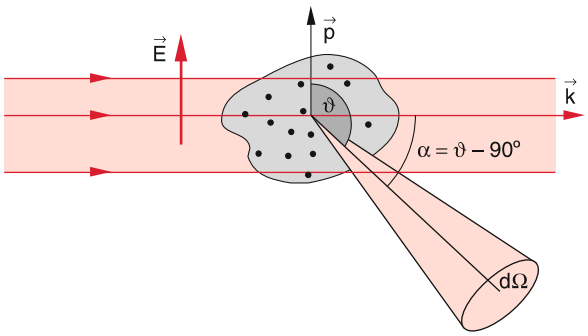


Fig. 10.64 Measurement of the light power $P_s(\vartheta)$, scattered by the angle ϑ against the electric field vector $\boldsymbol{E}$ into the solid angle dΩ

$$P_S(\vartheta) = \frac{Ne^2x_0^2\omega^4}{32\pi^2\varepsilon_0c^3}\sin^2\vartheta. \tag{10.98a}$$

Inserting for x_0^2 the expression (8.6b) we finally obtain

$$P_S(\omega,\vartheta) = \frac{Ne^4E_0^2\sin^2\vartheta}{32\pi^2m^2\varepsilon_0c^3}\cdot\frac{\omega^4}{(\omega_0^2-\omega^2)^2+\gamma^2\omega^2} \tag{10.98b}$$

The total power of the scattered radiation integrated over all angles ϑ and emitted into all directions ($\Omega = 4\pi$) is then

$$P_S(\omega) = \frac{Ne^4E_0^2}{12\pi\varepsilon_0m^2c^3}\cdot\frac{\omega^4}{(\omega_0^2-\omega^2)^2+\gamma^2\omega^2}, \tag{10.98c}$$

10.9.2 Scattering Cross Sections

We define the ratio

$$\sigma_S = (P_S/N)/I_i \tag{10.99}$$

of the power P_S/N scattered by one atom and the incident light intensity $I_i = 1/2\varepsilon_0cE_0^2$ as scattering cross section σ with the dimension [σ] = 1 m^2). This definition has the following descriptive meaning:

The scattering of light by an atom can be described by a circular disc with area σ. All the light passing through this area is completely scattered.

The time averaged radiation power scattered by N atoms is then

$$\overline{P}_S = N \cdot \sigma_S \cdot I_i.$$

From (10.98c) we get the scattering cross section for light scattering by atoms or molecules with a mean distance $d > \lambda$ (**Rayleigh Scattering**)

$$\sigma_S = \frac{e^4}{6\pi\varepsilon_0^2c^4m^2}\cdot\frac{\omega^4}{(\omega_0^2-\omega^2)^2+\omega^2\gamma^2}. \tag{10.100}$$

The scattering cross section takes particular large values around the resonance frequency ω_0, i.e. when the frequency of the incident light is nearly equal to the resonance frequency of the atoms or molecules (**Resonance–Rayleigh Scattering**).

The maximum σ_m of $\sigma(\omega)$ is at the frequency

$$\omega_m = \omega_0(1 - \gamma^2/2\omega_0^2)^{-1/2}, \qquad (10.101)$$

This follows from $d\sigma_S/d\omega_{\omega_m} = 0$ for $\omega = \omega_m$.

If the incident light is not monochromatic but has a spectral bandwidth $\Delta\omega$ with $\gamma < \Delta\omega \ll \omega_0$ the average of the scattering cross section is obtained by integration of (10, 100) over the frequency range $\Delta\omega$ [18]. For $\Delta\omega \ll \omega$ one obtains

$$\sigma_S(\omega) \propto \omega^4. \qquad (10.102)$$

This shows that the scattering cross section strongly increases with increasing frequency ω.

Some examples shall illustrate the different aspects of light scattering.

10.9.3 Scattering by Micro-particles; Mie-Scattering

When light is scattered not by atoms or molecules but by small solid micro-particles (dust, cigarette smoke, etc.) or by small liquid droplets (fog), a partial coherent scattering occurs if the diameter of the particles is small compared with the wavelength λ. In this case the phase differences between the partial waves scattered by the atoms of the particle are small compared with 2π ($\Delta\varphi \ll 2\pi$). This means that the amplitudes of the partial waves superimpose nearly with the same phase. The total intensity scattered by the particle is then

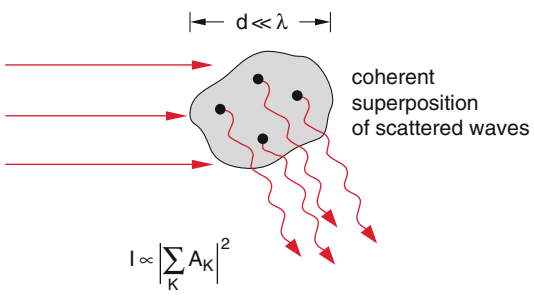


Fig. 10.65 Scattering of light by micro-particles with diameters $d \ll \lambda$ (Mie-Scattering)

$$I \propto \left|\sum_{K=1}^{N} A_K^2\right| \qquad (10.103)$$

where A_k is the scattering amplitude of the kth atom in the particle with N atoms (Fig. 10.65). Even if the atoms perform random motions with path lengths $s \ll \lambda$ this changes the phase differences only by $\Delta\varphi \ll 2\pi$. The scattered intensity then increases with d^6 as long as the diameter d of the particles is still small compared to the wavelength λ.

$$P_S \propto \left|\sum A_K\right|^2 = (N \cdot A_S)^2 = N^2 \cdot P_S(\text{Atom}) \qquad (10.104)$$

Example

A micro particle with $d = 0.05\,\mu m = 50\,nm$ consists of about $N = 10^6$ atoms. The light intensity scattered by these atoms at $\lambda = 500\,nm$ is about 10^6-times higher than for incoherent scattering by the different atoms.

When the diameters of the particles reach the wavelength λ the scattered intensity strongly depends on the diameter d, and also on the material of the particle and its surface quality. Now constructive as well as destructive interference between the different partial scattered waves can occur, depending on their optical path difference. The accurate theoretical treatment of this Mie-scattering (*Gustav Mie, 1886–1957*) demands significant mathematical efforts which exceed the framework of this introduction [14–17].

Interference, diffraction and scattering are responsible for many optical phenomena in our atmosphere. We will discuss this in the next section.

10.10 Optical Phenomena in Our Atmosphere

We start with atmospheric phenomena, which are based on light scattering.

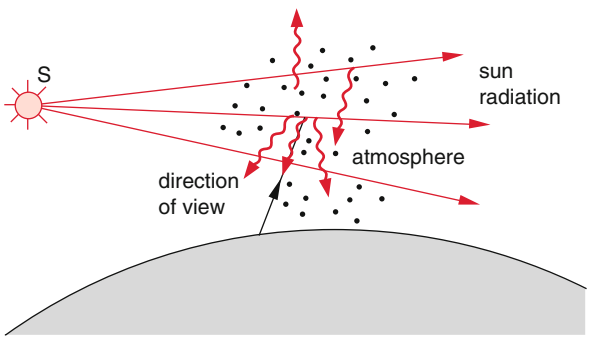


Fig. 10.66 Scattering of the sun light in the earth atmosphere

10.10.1 Light Scattering in Our Atmosphere

Even if we do not look towards our sun, we see on a sunny day a bright blue sky or white clouds. This is due to the scattering of the sun light by molecules, micro particles such as water droplets, dust particles or aerosols in the atmosphere (Fig. 10.66). For astronauts outside of our atmosphere the sky is dark (besides some bright stars) if they look away from the sun.

We will now discuss and answer the following questions:

10.10.1.1 Why Is the Unclouded Sky Blue?

The blue color of the sky is determined by three factors:

- The spectral intensity distribution $I(\lambda)$ of the sun radiation, which has a maximum at $\lambda = 455\,\text{nm}$ (Fig. 10.67). This is discussed in more detail in Vol. 3, Sect. 3.1.
- The wavelength-dependence of the scattering cross section $\sigma(\lambda)$, which varies with $1/\lambda^4$.
- The spectral distribution of the detection sensitivity $\eta(\lambda)$ of the human eye which is maximum at the wavelength λ_m (biological adaption).

The color of the sky as perceived be our eye, is therefore determined by the signal

$$S(\lambda) \propto I(\lambda) \cdot \sigma(\lambda) \cdot \eta(\lambda). \tag{10.105}$$

registered by our brain.

The absorption or emission wavelengths of the molecules in our atmosphere (N_2, O_2, H_2) are all in the ultraviolet region at $\lambda < 200$ nm. For the visible range the frequencies ω are therefore far away from the resonance frequency ω_0. This means that the term $\omega_0^2 - \omega^2$ in (10.100) is large compared to $\omega \cdot \gamma$ but does not vary much over the visible region. We can then set $\sigma(\omega) \propto \omega^4 \rightarrow \sigma(\lambda) \propto 1/\lambda^4$.

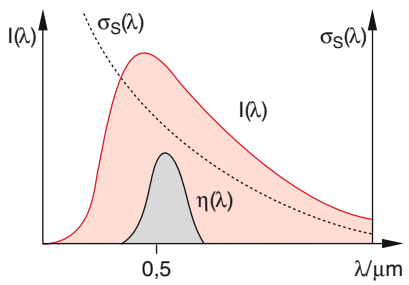


Fig. 10.67 Spectral intensity distribution $I(\lambda)$ of the sun radiation. Scattering cross section $\sigma_S(\lambda)$ and spectral dependence of the sensitivity $\eta(\lambda)$ of the human eye

Example

For $\omega = 1/3\omega_0$, $\gamma = 10^8\,\text{s}^{-1}$ $\omega_0 = 10^{15}\,\text{s}^{-1} \Rightarrow$ $\left(\omega_0^2 - \omega^2\right)^2 = 0.8\omega_0^4 \gg (\omega^2\gamma^2)$.

In Fig. 10.68 the scattering of the light beam of an argon laser is shown. Without scattering the laser beam would not be visible from the side. In order to determine the spectral maximum of $S(\lambda)$ we have to investigate the penetration depth L_i of the radiation with wavelength λ into the earth atmosphere. If the main contribution of the attenuation is due to Mie scattering by particles with scattering cross section σ_S we obtain

$$I(L) = I_0 \cdot \text{e}^{-\text{n}\cdot\sigma\cdot L}$$

and for the extinction length

$$L_e \approx \frac{1}{n \cdot \sigma_S}, \tag{10.106}$$

where n is the number density of scattering particles (number per unit volume). After the path length L_e has the intensity of the incident radiation decreased to 1/e of its initial value.

Other contributions to the attenuation are Rayleigh scattering and absorption by molecules in the atmosphere. The main gas components are N_2, O_2, O_3 and CO_2. The first two components do not absorb in the visible and UV region, O_3 which represents only a minor fraction of molecules, absorbs in the UV below 350 nm and protect us from the dangerous region of UV radiation of the sun, CO_2 absorbs in the infrared region. Therefore absorption plays only a minor role for the attenuation of the sun radiation. The Rayleigh scattering cross section is proportional to ω^4 or $1/\lambda^4$.

With $\sigma_S \propto 1/\lambda^4$ we get

$$L_e \propto \frac{\lambda^4}{n}. \tag{10.107}$$

Example

Typical cross sections for Rayleigh scattering by nitrogen molecules are $\sigma_S(\lambda_0) \approx 3 \times 10^{-31}\,\text{m}^2$ for $\lambda = 600\,\text{nm}$. With a density $n = 10^{25}\text{m}^{-3}$ of N_2-molecules at atmospheric pressure we get

$$L_e \approx 3 \times 10^5 \left(\frac{\lambda}{\lambda_0}\right)^4 \text{m}.$$

For $\lambda = 400\,\text{nm}$ (blue light) $\Rightarrow L_e = 60\,\text{km}$, for $\lambda = 700\,\text{nm}$ (red light) $\Rightarrow L_e = 550\,\text{km}$.

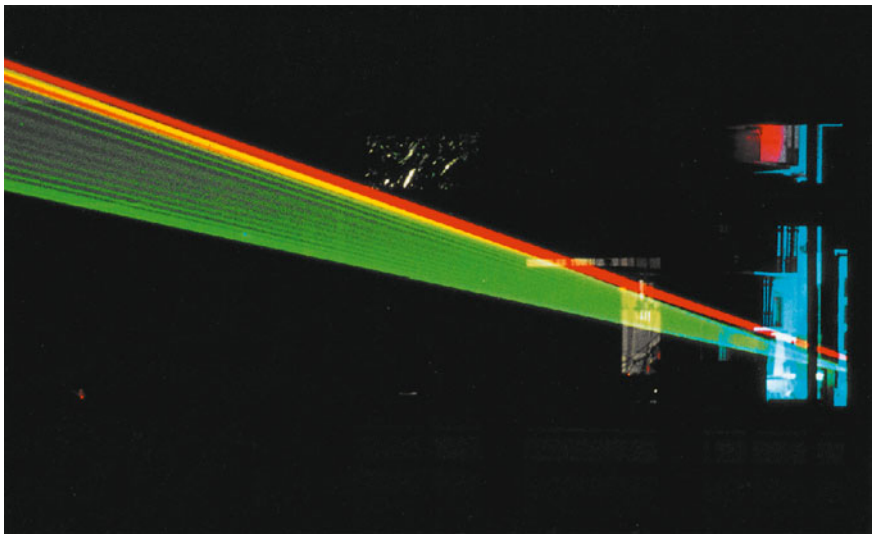


Fig. 10.68 Light scattering of the green output beam of an argon laser and the red beam of a krypton laser. The two beams are sent through the lab window (reflections) into the night sky. The yellow beam is a color mix of red and green-blue on the film. The beams are visible from the side only because of scattering (H. J. Foth, Kaiserslautern)

These numerical examples show that the attenuation of sun light by Rayleigh scattering only plays an essential role in the morning and evening, where the solar altitude is low and the path length of sun radiation through the atmosphere is large (Fig. 10.70). However, the contribution of Mie scattering due to dust and aerosol particles, water droplets and microscopic ice crystals is much higher than that of Rayleigh scattering. Therefore the attenuation of the blue contribution in the sun light is even at noon noticeable. On high mountains one observers a color of the sky that is shifted towards the UV region and appears to the eye as dark blue (Fig. 10.68).

Since the sun light is scattered into all directions, part of the sunlight, in particular the blue contribution, is scattered onto the earth surface. Therefore we see a blue sky when we do not look into the direction of the sun. Part of the sun light is scattered back into outer space. The earth therefore appears for an observer outside the earth (for instance from the moon) as the "blue planet", although part of this blue appearance is due to light scattered from the oceans.

Note The detailed treatment of light scattering in the atmosphere is very complex and can be found in [16–18].

10.10.1.2 Why Is the Sky Light Partially Polarized?

When one looks through a polarization filter into the blue sky one finds by rotating the filter, that the sky light is partially polarized. This can be explained as follows:

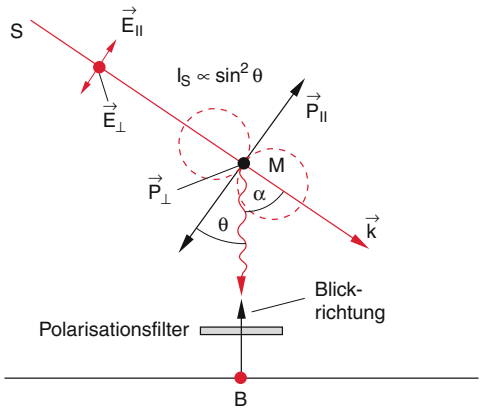


Fig. 10.69 Explanation of the partial polarization of the sun light scattered in the atmosphere

The molecular dipoles, induced by the sun light, oscillate in a plane perpendicular to the incidence direction $\boldsymbol{k}$ of the light (Fig. 10.69).

In this plane they have randomly orientated oscillation directions because the sun light is unpolarized. The dipoles oscillating in the plane SMB in Fig. 10.69 radiate into the direction towards the observer B the fraction $I_S = I_0 \cdot \sin^2\theta = I_0 \cdot \cos^2\alpha$ which is polarized in the plane SMB. For the dipoles oscillating perpendicular to the plane SMB the angle θ is $\theta = 90°$. This means their maximum scattering intensity if towards the observer B.

The component of the scattered radiation polarized perpendicular to the plane SMB is therefore stronger than the parallel component. The degree of polarization

$$PG = \frac{I_{\parallel} - I_{\perp}}{I_{\parallel} + I_{\perp}} = \frac{1 - \cos^2 \alpha}{1 + \cos^2 \alpha} \qquad (10.108)$$

depends on the angle $\alpha = 90° - \theta$ between the viewing direction *BM* and the direction *SM* of the sun radiation.

Bees use this partial polarization for their orientation.

10.10.1.3 Why Appears the Rising and Setting Sun Reddish?

This is also due to light scattering in the atmosphere. For the low position of the sun in the morning and evening the path length of the sun radiation through the atmosphere becomes very long. The observer looking into the direction to the sun observes the direct radiation which is attenuated by Rayleigh and Mie scattering (Fig. 10.70). The blue spectrum is much more scattered out of the incident direction than the red one. Therefore the spectral distribution has shifted to the red.

The sun radiation is also strongly attenuated on its way through the atmosphere. This effect is much more pronounced for low sun positions because the sun radiation now propagates a long way through the lower atmosphere, where besides the air molecules water droplets, ice-crystals and dust particles give a large contribution to the attenuation of the radiation. Therefore it is possible to look with the naked eye directly into the red sun. This would not be possible without filters at noon time, because the strong sun radiation, in particular the UV-fraction would damage the retina of the eye.

10.10.1.4 Why Appear Faraway Mountains Blue?

The sky light scattered by faraway mountains reach the observer after propagating through the low atmosphere (Fig. 10.70). Here the main contribution for the attenuation is Mie scattering by water droplets, dust particles and micro ice crystals. The size of these particles in the lower atmosphere is of the same order of magnitude as the wavelength λ. For such sizes of the scattering particles the Mie cross section does not strongly depend on the wavelength. Therefore the wavelength distribution of the radiation reaching the observer, is about the same as that, incident onto the mountains, which is the blue shifted radiation scattered by the sky.

For longer ways through the lower atmosphere the contribution of the Rayleigh scattering increases. This causes that the blue part of the spectrum is more strongly scattered out of the observation direction than the red part. This brings about a red shift of the observed radiation. Therefore very faraway mountains appear in a blue-white color [19].

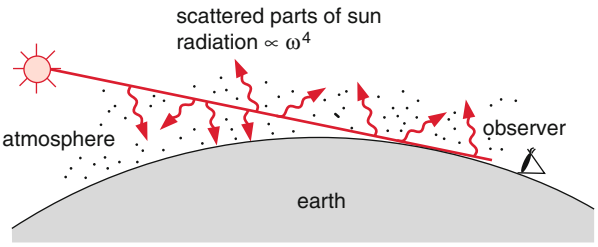


Fig. 10.70 Explanation of the reddish color of the rising or setting sun

10.10.2 Halo Phenomena

Under certain weather conditions one observes a colored ring around the sun, where the inner edge appears red and the outer edge blue (Fig. 10.71). This phenomenon is called *Halo* (= gloriole). Its formation is due, similar to the origin of the rainbow, to refraction and reflection of sun light. The refracting objects are here no spheres as for rainbows, but cylindrical ice crystals with hexagonal basis (Fig. 10.72) which are formed in the higher atmosphere. For a symmetrical ray path the minimum deflection angle is $\delta_{\min} = 22°$ for a refractive index $n = 1.31$ (see Sect. 9.4 and Problem 10.13).

Fig. 10.71 Halo around the sun (Andreas Möller - CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=44286865)

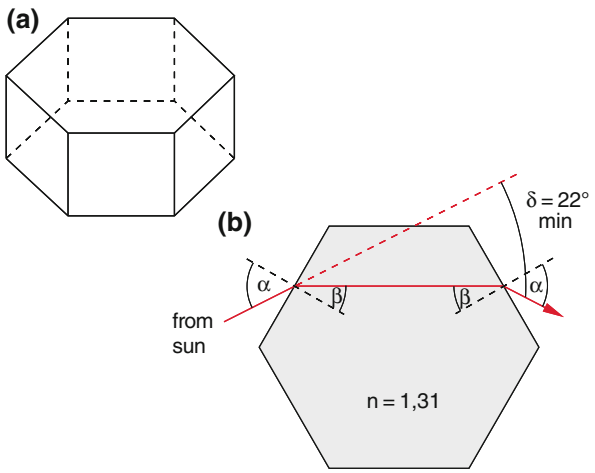


Fig. 10.72 Explanation of the halo. **a**) Rhombohedra ice-crystal with hexagonal basis, **b**) ray path for minimum deflection

Since the ice crystals in the atmosphere are randomly orientated, all possible angles α of incidence occur. For the case of minimum deflection is $d\delta/d\alpha = 0$. For δ_{min} therefore many angles of incidence in the interval $(\alpha_S \pm \Delta\alpha)$ contribute to the same deflection angle similar to the rainbow effect (Fig. 9.71). From all ice crystals those with the orientation that results in a symmetric ray path contributes most of the light deflected by 22°. Therefore at 22° appears an intensity maximum.

10.10.3 Aureole Around the Moon

Just before a bad weather period one can observe colored rings around the moon. They have, however, a reverse sequence $\lambda(r)$ of colors compared to the halo (Fig. 10.73). Therefore this *aureole* must be caused by another physical phenomenon as the *halo*. *Fraunhofer* recognized already in 1825 that the cause of this moon aureole is not refraction but diffraction by small water droplets or ice crystals in the atmosphere.

Since the central diffraction maximum for diffraction by a spherical droplet with a diameter d covers the angular range $\Delta\theta = \pm 1.2\lambda/d$ the diameter d of the droplets must be smaller than $d < 1.2\lambda/\Delta\theta_m$ in order to make the aureole larger than the angular diameter $\Delta\theta_M = 0.5° = 8.7 \times 10^{-3}$ rad of the moon disc. For λ = 500 nm this gives $d < 70\,\mu m$.

Often this colored ring with the blue inner edge and the red outer edge is called the Corona of the moon [20]. This name should be not confused with the *corona* of the sun, which is due to its very hot outer atmosphere and not to any diffraction effects. Its radiation can be seen during a sun eclipse.

Fig. 10.74 Glory and shadow of a plane, observed from the plane above the clouds

10.10.4 Glory Phenomena

When looking out of a plane which flies above the clouds, one can see (if the position of the sun is opposite to the viewing direction) the shadow of the plane surrounded by a bright colored circular disc (Fig. 10.74). Often a similar phenomenon can be observed when standing on the top of a mountain above the clouds with the backside towards the sun. One can then see its own shadow, where the head is

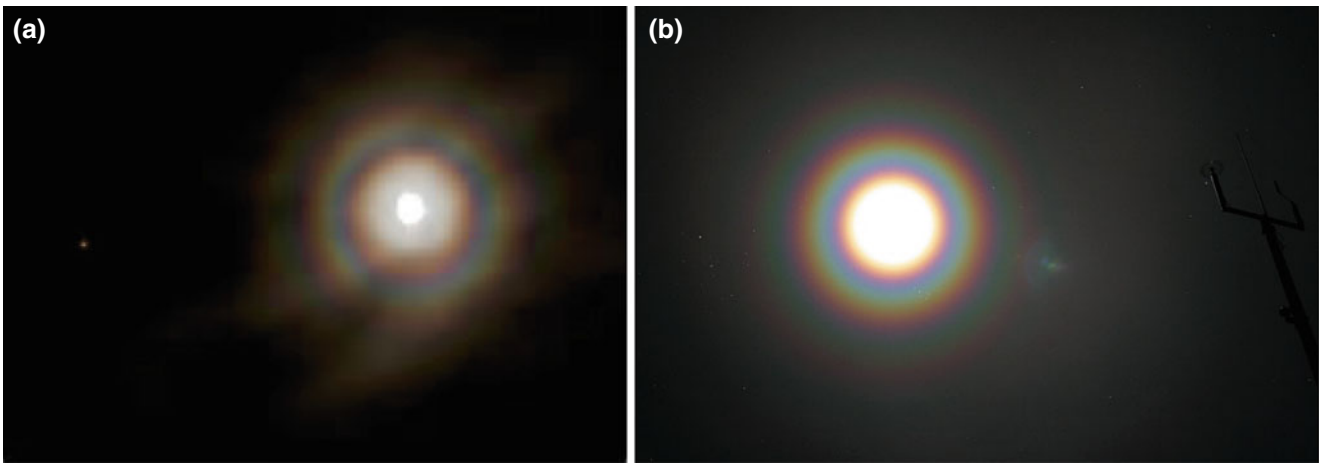


Fig. 10.73 **a**) Aureole around the moon [Benjamin Kühne], **b**) around the sun [Karl Kaiser]

surrounded by a glory like the gloriole around the heads of the Saints on ancient paintings.

Its physical origin is more difficult to explain than the rainbow or the aureole. Only recently it was recognized [21] that it is based on the quantum-mechanical tunnel effect (see Vol. 3, Sect. 4.2.3). This is illustrated in Fig. 10.75.

The incident sunlight partly penetrates into a water droplet and is then refracted and reflected at the spherical surface of the droplet. It turns out, however, that this alone cannot correctly describe the observed phenomena. Even light that falls closely above or below the droplet can tunnel through the droplet surface into the inside (Fig. 10.75) and propagates by total reflection many times along the inner part of the surface [20]. At each reflection part of the wave can tunnel back out of the droplet. If the exit location can be seen by the observer, he sees the blaze of glory. The light is deflected in all directions, different from the situation for the rainbow.

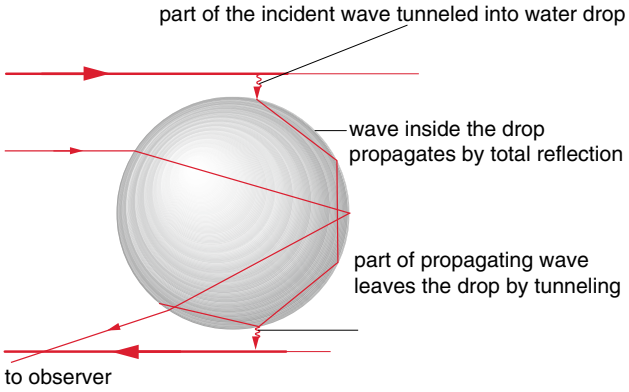


Fig. 10.75 Explanation of the glory phenomenon

Summary

- Interference phenomena can be observed when two or more coherent partial waves with locally dependent phase differences are superimposed in a defined space region. The coherence volume is the maximum volume where still coherent superposition is possible.
- The coherent partial waves can be realized either by phase coupling of two or more radiation sources or by splitting of one wave into two or more partial waves, which are again superimposed after traversing different path lengths Δs. This superposition gives maximum intensity, if the path differences $\Delta s = m \cdot \lambda$ are an integer multiple of the wavelength.
- *Albert Abraham Michelson* could experimentally prove, using a two-beam interferometer, that the speed of light is independent of the motion of source or observer.
- Multiple-beam interference is used in the Fabry-Perot-interferometer (FPI) for the precise measurement of optical wavelengths. For dielectric multi-layer mirrors it allows the realization of any requested wavelength-dependent reflectivity $R(\lambda)$.
- The spatial propagation of waves can be described by Huygens's principle which states that each point of a phase surface can be regarded as the source of a spherical wave (secondary wave). The total wave is then the superposition of all secondary waves.
- The diffraction of waves can be taken as interference of secondary waves which are emitted from a spatially confined region.
- The angular intensity distribution of a plane wave diffracted by a slit with width b can be described as

$$I(\theta) = I_0 \frac{\sin^2[\pi(b/\lambda)\sin\theta]}{[\pi(b/\lambda)\sin\theta]^2},$$

  where θ is the angle against the direction of the incident wave.
- The rotational symmetric intensity distribution of a plane wave diffracted by a circular aperture with radius R is

$$I(\theta) = I_0 \frac{J_1^2[2\pi(R/\lambda)\sin\theta]}{[2\pi(R/\lambda)\sin\theta]^2},$$

  where J_1 is the first order Bessel function.
- The intensity distribution of a diffraction grating is determined by two factors: (1) the diffraction by a single slit, (2) the interference of the partial waves transmitted by the different slits of the grating.
- Fraunhofer-diffraction describes the diffraction of parallel light beams, Fresnel diffraction that of divergent or convergent light. Fraunhofer-diffraction is observed in the far field at the distance $(z \gg b^2/\lambda)$ behind the diffracting aperture with diameter b, Fresnel diffraction is observed in the near field where still $z \gg b$ but where only a few Fresnel zones contribute to the field amplitude of the wave in the observation plane.
- Blocking the first Fresnel zone increases the intensity in the observation plane.
- With Fresnel zone plates (Fresnel lenses) the optical imaging of a light source can be realized by blocking either all even Fresnel zones or all odd ones. This yields constructive interference of all transmitted light beams and increases the intensity in the observation plane.
- Babinet's theorem states that two complementary areas where transparent and opaque areas are interchanged show the same diffraction structures (outside of regions described b geometrical optic).
- The amplitude distribution $E(x', y')$ of the Fraunhofer diffraction pattern is proportional to the Fourier-transform of the field distribution $E(x, y)$ in the object plane.
- The Fourier transform of a constant field amplitude within a rectangular opening $a \cdot b$ yields in the observation plane x', y' the diffraction pattern of two orthogonal infinitely extended slits with widths a and b resp.
- Light is scattered by atoms, molecules and microparticles. Coherent scattering occurs if there are temporally constant distances $b < \lambda$ between the different scattering centers. If these distances vary randomly in time, incoherent scattering is observed.
- For coherent scattering the scattered intensity is obtained by adding the amplitudes from the different scattering centers and then squaring the sum i.e. $I = (\Sigma A_k)^2$.
- For incoherent scattering the intensities of the different scattering centers are added: $I = \Sigma I_k$.
- The circular halo around the sun with an angular diameter of $2 \times 22^\circ$ is generated by refraction and reflection of light by hexagonal ice crystals in the higher atmosphere.
- The aureole around the moon is due to diffraction of the moon light by water droplets or ice crystals in the atmosphere.

Problems

10.1 (a) Show that the expression (10.5) describes for constant values of Δs hyperbolas $(x^2/a^2) - (y^2/b^2) = 1$. How depends a and b from Δs and from the distance $2d$ of the two virtual light sources?
(b) Calculate for $z_0 \gg d$ the distance between the vertexes of the two hyperbolas.

10.2 How large are the radii of the interference rings in the observation plane behind a Michelson interferometer that is illuminated by divergent light as a function of the path difference Δs?

10.3 Why is an interference pattern of parallel stripes observed behind a Michelson interferometer, when one of the mirrors M_1 or M_2 is slightly tilted?

10.4 What is the reflectivity R of a dielectric mirror for vertical incidence
(a) for one layer $n_H \cdot d = \lambda/4$
(b) $n_H \cdot d = \lambda/2$
(c) For (H, L) alternating layers consisting of two $\lambda/4$ layers with $n_H = 1.8$, $n_L = 1.3$ on a glass substrate with $n_S = 1.5$ in air with $n_0 = 1$?

10.5 Determine the intensity distribution of diffracted light behind a slit with width D when a parallel light beam with wavelength λ under the angle α_0 against the surface normal incides onto the slit. Show that the distribution $I(\alpha_0, \alpha)$ reduces to (10.43) for $\alpha_0 = 0$.

10.6 A parallel light beam with $\lambda = 480$ nm hits an optical grating with 1000 grooves per mm under the incidence angle $\alpha = 30°$ against the grating normal.
(a) At which angle β appears the first diffraction order? Does the second order exist?
(b) How large must be the blaze angle θ?
(c) How large is the difference $\Delta\beta$ for the two wavelengths $\lambda_1 = 480$ nm and $\lambda_2 = 481$ nm?
(d) What is the maximum entrance slit width b in a grating monochromator with a 100×100 mm^2 grating and focal lengths $f_1 = f_2 = 1$ m when the two wavelengths should be resolved? How large is the diffraction limited width of the entrance slit image?
(e) Under which angle α must the incident beam hit the grating when the diffracted light should be reflected back into the incidence direction (Littrow grating)?

10.7 An oil layer ($n = 1.6$) on a water surface is illuminated by light with $\lambda = 500$ nm It reflects maximum intensity when illuminated under the incidence angle $\alpha = 45°$. How thick is the layer? Which wavelength would be preferentially reflected for vertical incidence ($\alpha = 0$)?

10.8 Two plane parallel glass plates are placed on top of each other. At one edge a thin paper strip is placed between the plates, causing a wedge-shaped air layer between the plates. For vertical illumination with parallel light at the wavelength $\lambda = 589$ nm one observes 12 interference stripes per cm. What is the wedge angle?

10.9 The first slit in Young's double slit experiment may be twice as broad as the second slit. How does the intensity distribution looks like on a screen far behind the slits?

10.10 The first order diffraction maximum is located not right in the middle between the first and the second diffraction minimum. How large is the deviation from the middle point?

10.11 The diameter of a laser beam ($\lambda = 600$ nm) is enlarged by a telescope to a parallel beam with 1 m diameter d and is sent to the moon.
(a) How large is the light spot on the moon ($D = 380.000$ km), if air turbulence in the earth atmosphere can be neglected?
(b) Which power of the light reflected by the retroreflector on the moon ($0.5 \times 0.5\,m^2$ area) reaches the telescope, when 10^8 W have been sent through the telescope to the moon?
(c) How large would this power be, if the light is diffusively and uniformly reflected by the moon surface with the reflectivity $R = 0.3$ into all directions within the solid angle $\Omega = 2\pi$ (without retroreflector)?

10.12 (a) Show, that for a single anti-reflection-layer (10.37a, 10.37b) is valid. Take into account the two possibilities for the refractive index n_1 of the layer and select the correct thickness d of the layer.
(b) Show that a satisfactory result can be obtained even for only two reflected rays.

10.13 Show that the minimum deflection angle for the refraction by a hexagonal ice crystal with $n = 1.31$ is given by $\delta_{min} = 22°$.

10.14 Calculate the optical frequency ω_m for which the scattering cross section for light scattering becomes maximum. Compare the result with the frequency-dependent energy consumption of a damped forced oscillator (Vol. 1, Chap. 11).

References

1. L. Mandel, E. Wolf: Coherence properties of optical fields. Rev. Modern Physics **37**, 231 (1965)
2. R. Castell, W. Demtröder, A.Fischer, R. Kullmer H. Weickenmeier, K. Wickert: The accuracy of laser wavelength meters. Appl. Physics **B 38**, 1–10 (1985)
3. W. Demtröder. Laser Spectroscopy 5th ed. (Springer 2014)
4. A. Michelson:Experimental determinatgion of the velocity of light. Am. J. Science Series 3 Vol. **18**, 310 (1879)
5. B. Jaffe: Michelson and the speed of light. (Greenwood Press Westport 1979)
6. J.M. Vaughan: The Fabry-Perot Interferometer (Hilger, Bristol, 1989)
7. A. Thelen: Design of Optical Interference Coatings (McGraw Hill New York 1988)
8. A. Musset, A.Thelen: Multilayer Antireflection Coatings in: Progress in Optics **Vol. 3**, 8North Holland Amsterdam 1970)
9. W. Osten, E. Novak: Interferometry: Applications (SPIE Int Society for Optical Engineering 2004)
10. M.C. Hutley: Diffraction Gratings (Academic Press, 1982)
11. http://www.x-ray-optics.de/index.php/en/types-of-optics/diffracting-optics/fresnel-zone-plates
12. J. Mazumder, Aravindar Kar: Theory and Application of Laser Chemical Vapor Deposition. (Springer, US 1995)
13. G.B. Thoma, Jr.; R.L. Finney, (1996), *Calculus and Analytic Geometry* (9th ed.), Addison Wesley
14. M. Kerker: The Scattering of Light (Academic Press 1969)
15. C.F. Bohren, D.R. Hufmann: Light Scattering by small particles (Wiley, New York 1983)
16. D.L. Lockwood: Rayleigh and Mie-Scattering (Encyclopedia of Color Science and Technology (Springer Heidelberg 2015). https://link.springer.com/referenceworkentry/10.1007/978-3-642-27851-8_218-1
17. https://www.sciencedirect.com/topics/earth-and-planetary-sciences/mie-scattering
18. C.M. Sorensen,D.J. Fishbach: Patterns in Mie-Scattering. Optics Communication 173, 145 (2000)
19. H.J. Schlichting: Rote Sonne, blaue Berge. Physik in uns. Zeit **36**, 291 (2005)
20. H.M. Nussenzweig, Ligth Tunneling in Colouds. Appl. Optics **42**, 1588 (2003)

Optical Instruments

11

Our ability to see is probable the most important communication between the human individual and his surroundings. Although the human eye is, regarded from the optical standpoint, a lousy lens with many lens aberrations, it forms in combination with our brain, which corrects most of the aberrations, an admirable optical instrument, which can optimally adapt to the actual optical conditions.

Nevertheless it needs for many situations additional instruments, which can enlarge the perception range. They can increase the spatial resolution capacity (magnifying glass, microscope), the light intensity reaching the eye from weak sources (telescope) or broaden the accessible spectral range (image converter).

In this chapter we will represent the most important optical instruments, their advantages and their limitations. Furthermore the spectrograph and the monochromator, which are essential instruments for spectroscopy, are introduced. Their spectral resolution is compared with that of interferometers, which have been already discussed in Chap. 10.

11.1 The Human Eye

The human eye can be regarded as an adaptive optical instrument that can be optimized for different distances of the observed object and for different incident light intensities. Its biological composition is accordingly complex.

11.1.1 The Bio-physical Structure of the Eye

One distinguishes between the external eye (eyelid with eyelashes, lacrimal glands, eye muscles), the eyeball as the main part of the eye lens, and the retina with the optic nerves as detector (Fig. 11.1).

The eye ball is nearly spherical with a diameter of about 22 mm. It is enclosed by the opaque white sclera S which is connected on the front side with the transparent bulged cornea C. behind the cornea is the iris I which has a circular aperture with variable diameter (pupil P) which can adapt to the incident light intensity (controlled by the brain). The space between cornea C and iris I is the anterior chamber of the eye, which is filled with a transparent diluted liquid. Behind the iris is the biconvex eye-lens which consists of many transparent layers. Its radius of curvature is controlled by the eye muscle M which adapts the focal length of the lens according to the distance of the observed object (accommodation). The focal length of the exe is however not only determined by the lens, but also by the cornea, the eye chamber liquid and the vitreous body of the eye ball. Since the outer interface of the cornea faces atmospheric air, but the inner interface the liquid of the anterior eye chamber the focal length f_1 of the object side is different from the focal length f_2 on the image side (Fig. 11.2).

For the discussion of the optical imaging of the human eye, it can be replaced by a lens with variable focal length. For observed objects in an infinite distance d (relaxed eye) is $f_1 = 17\,\text{mm}$ and $f_2 = 22\,\text{mm}$. For observed objects in a near distance d down to d = 10 cm (visual range) the eye lens must be stronger curved by the eye muscle and f_1 decreases to $f_1 = 14$ mm and $f_2 = 19$ mm.

The light sensitive part of the eye is the retina which consists of several layers (Fig. 11.3). At first (seen from the eye lens) a layer of nerve fibers covers a layer of ganglia and bipolar cells followed by a layer where the actual photoreceptors (rods and cones) are located. The total retina has much more rods (120 Millions) than cones (10 Millions). Only in the area of the sharpest seeing (fovea) where most of the light at normal seeing is focused, are exclusively cones with a density of $14.000/\text{mm}^2$, which decreases strongly towards the edge of the retina.

The rods are more sensitive than the cones, but they are color-blind, i.e. they can only distinguish between dark and bright contrary to the cones which exist as three types with different types of receptors for red, green and blue. In Fig. 11.4 the relative sensitivity $\eta(\lambda)/\eta_m$ is depicted for the

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_11

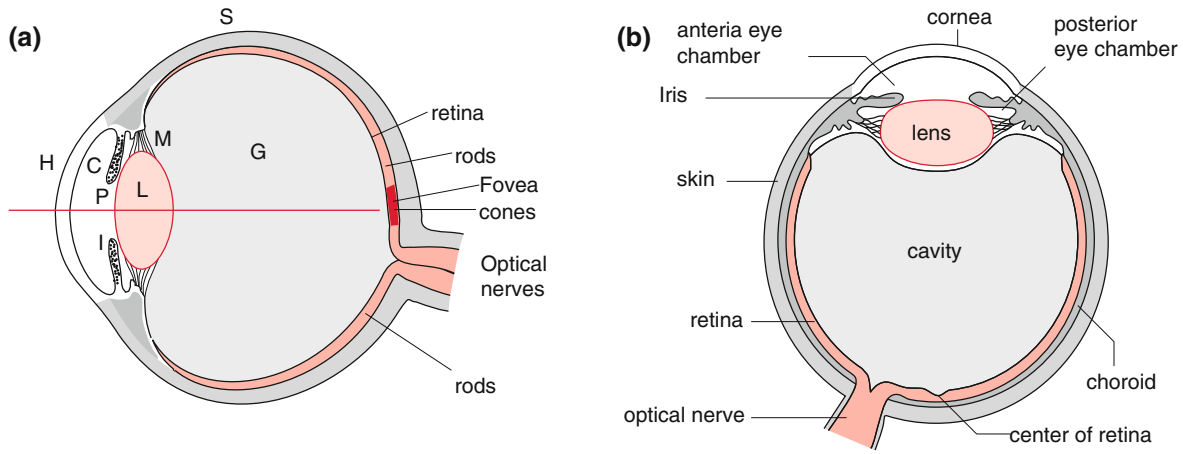


Fig. 11.1 **a**) Vertical cut through the human eye. **b**) Horizontal cut through the right eye seen from above

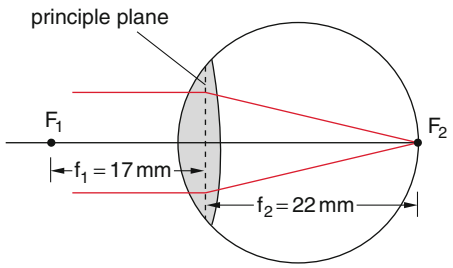


Fig. 11.2 Compensation representation of the eye by a lens with focal length f_1 on the object side and f_2 on the image side

different types of cones, where η_m is the maximum sensitivity of cone type b. At sufficiently high light intensity we see only with the cones, at darkness only with the rods, at twilight with both. Since the rods are more sensitive than the cones it is difficult to distinguish colors at twilight [1, 2]. Recently new cells (ganglion cells) have been discovered in the retina, which control the daily rhythm of our body. When these cells receive light, they alter their electrical conductivity. This generates neuronal signals that are transferred to the brain.

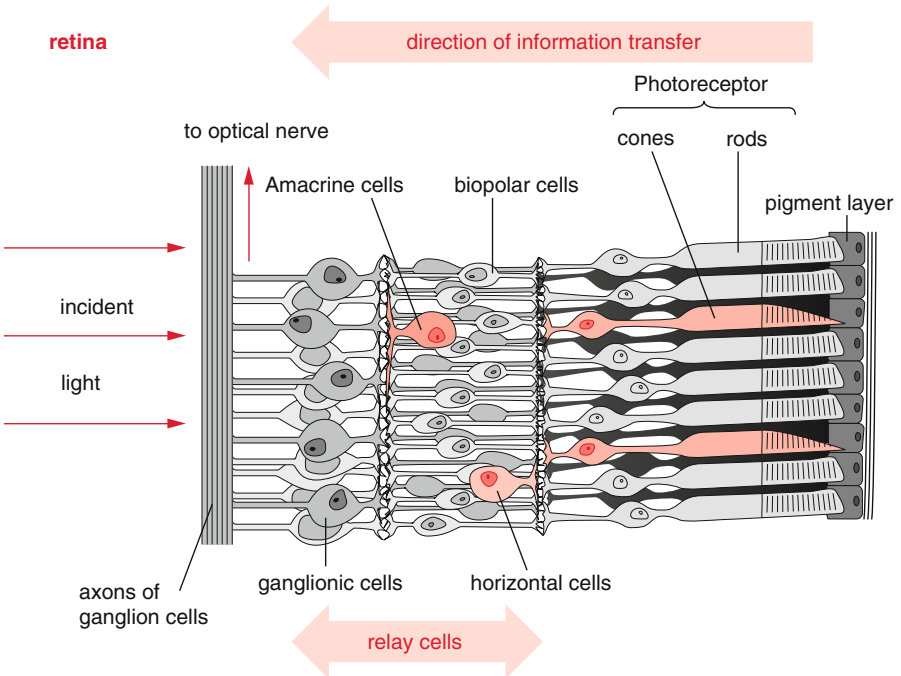


Fig. 11.3 Detailed structure of the retina [wikipedia org/wiki/netzhaut]

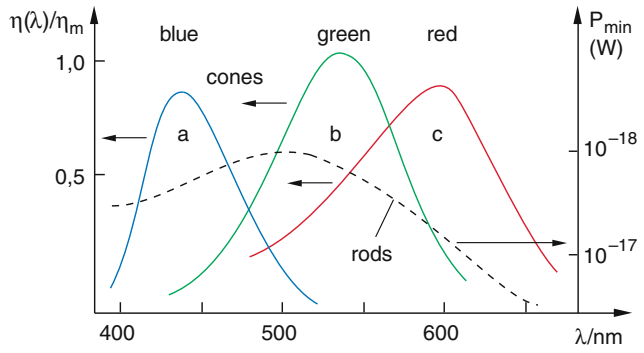


Fig. 11.4 Relative spectral sensitivity of the three receptor cells *a*, *b* and *c* in the cones and of the rhodopsin pigments in the rods (dashed curve) The right ordinate scale gives the minimum power, incident onto the retina, which can be still detected by the cones, which is smaller than that for the rods

It is interesting that the incident light has to pass through all layers of the retina before it reaches the photoreceptors. This is certainly not optimized with respect to the function of the eye as optical imaging system. The electrical output signals of the photoreceptors has to be sent back to the nerve cells in the first layer of the retina, which then send their output signals to the brain. The question arises whether the light on its way through all layers of the retina is not scattered, which would blur the image on the photoreceptors and therefore would decrease the spatial resolution. Only recently detailed investigations have shown, that some cells in the retina act as optical fibers which guide the light to the photoreceptors without scattering.

11.1.2 Short- and Far-Sightedness

For a short-sighted eye the focal length f_2 on the image side is too small. The eye muscle cannot sufficiently stretch the eye lens (for instance if the eye socket is too small). The radius of curvature of the lens is then too small. For all objects at a far distance the image is produced before the retina (Fig. 11.5a), while objects at very small distances are correctly imaged. Shortsightedness can be corrected by an additional diverging lens (Fig. 11.5a) which can be either carried as eyeglasses or as contact lenses.

For a farsighted eye the eye lens cannot be sufficiently contorted (for instance because of the visual fatigue of the eye muscle for older people, called *presbyopia*). Therefore the focal length of the image side is too large and the image of objects lies behind the retina. Here a converging lens is needed for correction (Fig. 11.5b)

Since the eye acts as an imaging lens, all lens aberrations, discussed in Sect. 9.5.5 (for instance astigmatism) can occur. They can be corrected by special grounded eye glasses. For astigmatism these are a combination of spherical and cylindrical lenses [3].

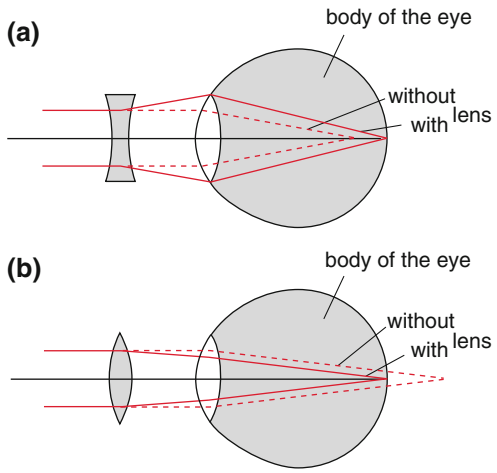


Fig. 11.5 **a**) the short-sighted eye with and without a diverging lens, **b**) the far sighted eye with and without a converging lens

11.1.3 Spatial Resolution and Sensitivity of the Eye

The nearer an object is brought to the eye the larger appears its size, i.e. the larger becomes the angle ε between the light rays from the edges of the object (Fig. 11.6). At a distance s of the object with diameter G we get for the visual angle ε

$$\tan \varepsilon/2 = \frac{1}{2}\frac{G}{s} \Rightarrow \varepsilon \approx \frac{G}{s} \tag{11.1}$$

An object at the distance g has an image distance b according to the lens equation

$$\frac{f_1}{g} + \frac{f_2}{b} = 1 \tag{11.2}$$

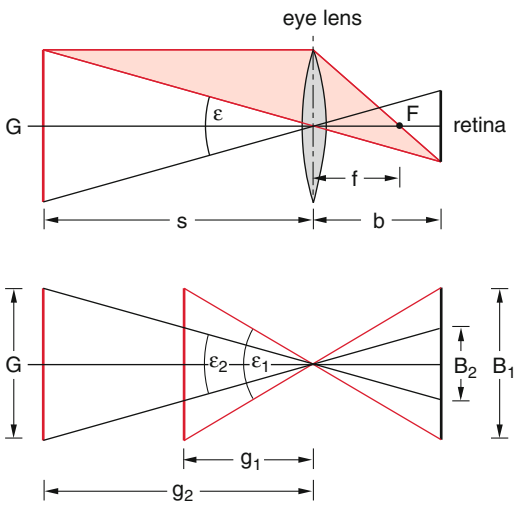


Fig. 11.6 Definition of the visual angle

Note: Equation (11.2) differs from (9.26) because the media before (air) and behind (eye liquid) are different and therefore $f_1 \neq f_2$. It can be derived in a similar way as (9.26) following the discussion in Sect. 9.5.2 (see problem 11.3 and [4]).

Since the distance b between eye lens and retina is fixed by the geometry of the eye the focal length of the eye lens has to be adapted to the distance g of the object by changing the curvature of the eye lens in such a way that the image occurs exactly on the retina. This is, however, only possible down to a minimum distance s_{min} of the object, which differs for different persons but lies always around $s_{min} = 10$ cm. To avoid fast fatigue of the eye the distance of objects should not be smaller than $s_0 = 25$ cm. This distance s_0 is called the *clear visual range* and the corresponding visual angle is ε_0.

Example

An object at a distance $s = 1$ m can be imaged onto the retina with distance $b = 22$ mm by the eye lens with $f_1 = 16$ mm if the focal length f_2 becomes $f_2 = 21.6$ mm. For $s = 15$ cm and $f_1 = 14$ mm, $b = 22$ mm is $f_2 = 19.95$ mm because the image distance is for both cases nearly the same. Without the change of $f_2 = 21.6$ mm the focal length on the object side had to be $f_1 = 2.7$ mm, which is impossible because of the geometry of the eye. The change of both focal lengths is therefore an optimization process where the largest possible range for the focus depth can be reached at a minimum change of the curvature of the eye lens.

The smallest still resolvable visual angle ε_{min} is determined by two factors:

(1) The diffraction by the eye pupil (see Sect. 11.3)
(2) The mutual distance between the photoreceptors on the retina.

Nature has optimized this distance in such a way that both limitations become equal. They limit the minimum still resolvable visual angle to $\varepsilon_{min} = 0.00028\text{rad} \leftrightarrow 1'$. This implies that two object points in a plane at the clear visual range $s_0 = 25$ cm with a mutual distance smaller than

$$\Delta x_{min} \approx s_0 \cdot \varepsilon_{min} \approx 25 \cdot 2.8 \cdot 10^{-4}\,\text{cm} = 73\,\mu\text{m}$$

cannot be resolved, i.e. they cannot be recognizes as two different objects.

Example

Many older printers operate with a spatial resolution of 360 dpi (dots per inch). This corresponds to a distance between two points of 70 μm. If the printed page is hold closer than 25 cm in front of the eye one can still see structures of the print letters. The present book is printed with a resolution of 2540 dpi!

The sensitivity of the human eye for the detection of very small light intensities is astonishing. For eyes adapted to darkness the brain can still perceive signals from the rods and cones in the retina if the received light power is as low as 10^{-17} W, whereas the maximum light power which can be received without damage of the retina is 10^{-6} W. This illustrates the large range of light powers that can be handled by our eyes.

The light perception of our eyes is proportional to the logarithm of the incident light power, but also depends on the intensity received beforehand. The eye adapted to lightness stores the incident light power for about 50 μs, the eye adapted to darkness for about 500 μs. The determination of absolute light intensities from our visual light perception is therefore not reliable, whereas the relative comparison between the intensities reflected from two illuminated surfaces is very sensitive and reliable.

11.2 Magnifying Optical Instruments

The purpose of magnifying optical instruments is the magnification of the visual angle ε without coming below the clear visual range s_0. The angular magnification V of the instrument is defined as the ratio

$$V = \frac{\text{visual angle } \varepsilon \text{ with instrument}}{\text{visual angle } \varepsilon_0 \text{ without instrument}}$$

Magnifying instruments allow the resolution of finer details of the observed object which could not be resolved by the naked eye if the visual angle ε_0 of an object at the clear visual range s_0 is smaller than $\varepsilon_{min} = \Delta x_{min}/s_o = 1'$ (see Sect. 11.1.3).

Note: The angular magnification is generally not the same as the image ratio B/G, defined as the ratio of image size B divided by the object size G.

Since the optical instruments generally have a fixed focal length f, they can generate sharp images of object points A only in a fixed image plane $z = g$. If the object point A is

shifted by the distance Δz, the image in the plane $z = g$ becomes blurred out into a disc. The maximum shift Δz_{max} which still gives an image size of an object point below the resolution of the eye at the distance s_0 is called the ***focus depth*** of the optical instrument. The images within the focus depth are still regarded as sharp images. The focus depth depends on the diameter D_a of the aperture in the instrument. This can be seen as follows:

In Fig. 11.7 the point A is imaged by a lens into the point B. If the object A is shifted to A_f at the front side of the focus depth the image is a circle with diameter u. The same is true, if A is shifted to A_b at the backside of the focus depth. According to the theorem of intercepting lines we get from Fig. 11.7 the relation

$$\frac{u}{D_B} = \frac{b_b - b_0}{b_b}.$$

in a similar way the imaging of A_b gives the relation

$$\frac{u}{D_B} = \frac{b_0 - b_h}{b_h}.$$

using the lens equation

$$\frac{1}{a} + \frac{1}{b} = \frac{1}{f}$$

we obtain for the imaging of the points A_f, A and A_b the front focus depth

$$\Delta a_f = a_0 - a_f = \frac{b_0 f^2 u}{(b_0 - f)(D_B b_0 - D_B f + uf)} \tag{11.3a}$$

and for the back focus depth

$$\Delta a_b = a_b - a_0 = \frac{b_0 f^2 u}{(b_0 - f)(D_B b_0 - D_B f - uf)}. \tag{11.3b}$$

The front focus depth Δa_f is therefore smaller than the back focus depth Δa_b. Both are proportional to the square of the focal length f and increase with decreasing diameter D_B of the aperture. The focus depth is therefore increased by using a smaller aperture diameter D_a.

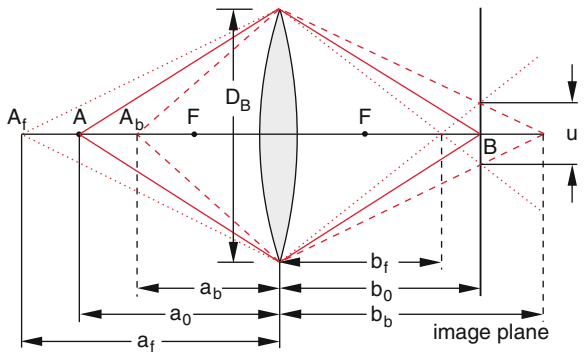


Fig. 11.7 Definition of the focus depth for imaging by a lens

Example

a_0 = 1 m, f = 50 mm, b_0 = 52.6 mm, u = 0.1 mm.

(a) For D_B = 1 cm => Δa_b = 0.24 m and Δa_f = 0.16 m.

(b) For D_a = 0.3 cm => Δa_b = 1.8 m and Δa_f = 0.40 m.

For case (a) the focus depth ranges from 1.24 to 0.84 m, for case (b) from 2.8 to 0.60 m.

11.2.1 Magnifying Glass

A magnifying glass is a lens with short focal length f. It is placed between object and eye in such a sway that the object lies in the focal plane of the lens (Fig. 11.8). Therefore a parallel light beam enters the eye and the object appears at infinite distance. The relaxed eye can adapt to this infinite distance. The object with diameter A appears for the eye under the angle $\varepsilon = A/f$. Without magnifying glass the object would appear at the clear visual range s_0 under the angle $\varepsilon_0 = A/s_0$.

The angular magnification of the magnifying glass is then

$$V_M = \frac{\tan\varepsilon}{\tan\varepsilon_0} = \frac{A}{f} \cdot \frac{s_0}{A} = \frac{s_0}{f}. \tag{11.3c}$$

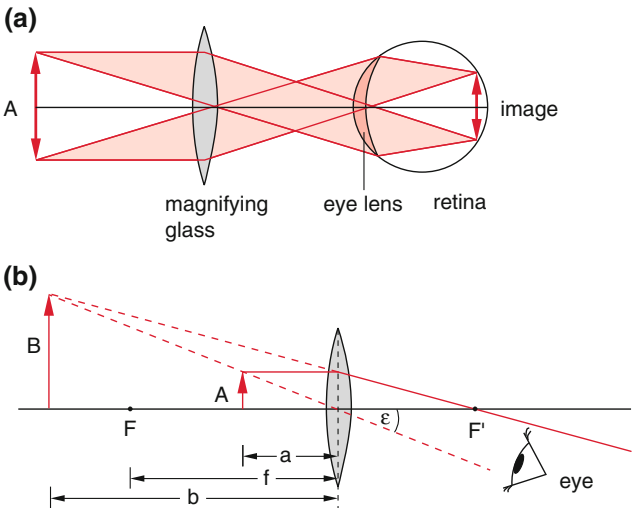


Fig. 11.8 Magnification of the visual angle ε by a magnifying glass. **a**) If the object A lies in the focal plane, **b**) if a < f

The angular magnification of the magnifying glass is equal to the ratio of clear visual range s_0 to the focal length f.

Example

$f = 2$ cm, $s_0 = 25$ cm => $V_M = 12.5$

The reason for the magnification is the small focal length f which allows one to bring the object much closer to the eye than the clear visual range s_0 where the eye sees the object at infinite distance and must therefore not accommodate to small distances but can completely relax.

The magnification can be further enlarged if the object is brought to the lens closer than the focal length ($s < f$). The object then does not appear at infinite distance but at the distance b (Fig. 11.8b) as a virtual image of A. (Fig. 11.9)

The magnification V_L then becomes with $B/b = A/a$ (theorem of intersecting lines)

$$V_L = \frac{\tan\varepsilon}{\tan\varepsilon_0} = \frac{B/b}{A/s_0} = \frac{A/g}{A/s_0} = \frac{s_0}{g}.$$

With the lens equation

we get

$$\frac{1}{f} = \frac{1}{g} + \frac{1}{b} \Rightarrow \frac{1}{g} = \frac{b-f}{b\cdot f}$$
$$\Rightarrow V_L = \frac{s_0(b-f)}{b\cdot f}.$$

For the clear visual range $b = -s_0$ we obtain

$$\Rightarrow V_L = \frac{s_0 + f}{f} = \frac{s_0}{f} + 1. \qquad (11.4)$$

The minimum object distance is

$$g_{min} = \frac{s_0 \cdot f}{s_0 + f}.$$

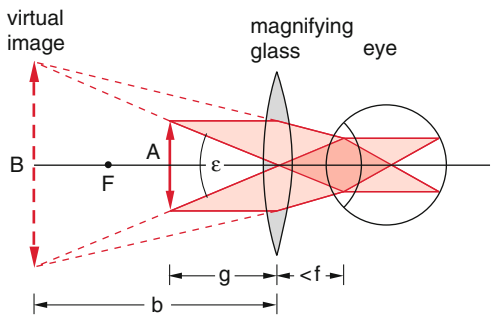


Fig. 11.9 Illustration of Eq. (11.4)

Example

$s_0 = 25$ cm, $f = 2$ cm => $g_{min} = 1.85$ cm => $V_L = 13.3$.

The eye lens now has to curve more strongly in order to focus the divergent light rays onto the retina.

11.2.2 The Microscope

The microscope allows a much larger magnification than the magnifying glass. Its basic design consists of two lenses (Fig. 11.10). The first lens (objective) generates a real intermediate image of the object in the focal plane of the second lens (ocular). Therefore again parallel light beams from each object point reach the eye similar to the situation for the magnifying glass in Fig. 11.8.

One can derive from Fig. 11.10, using the theorem of intercepting lines the relation $B/A = b/a$. With the lens equation for L_1 one obtains

$$\frac{1}{f_1} = \frac{1}{g} + \frac{1}{b} \quad \Rightarrow \quad b = \frac{g\cdot f_1}{g - f_1} = \frac{gf_1}{\delta}. \qquad (11.4a)$$

When the object is placed close to the focal plane of L_1 which gives $g = f_1 + \delta$ ($\delta << f_1$) we get $b >> g \Rightarrow B >> G$,

The ocular L_2 acts as magnifying glass for the intermediate image. It is

$$\tan\varepsilon = B_1/f_2 = \frac{G\cdot b}{g\cdot f_2}. \qquad (11.4b)$$

Without microscope the visual angle for a distance s_0 of the object would be

$$\tan\varepsilon_0 = \frac{G}{s_0}. \qquad (11.4c)$$

The angular magnification of the microscope is then

$$V_M = \frac{Gbs_0}{Ggf_2} = \frac{bs_0}{gf_2}. \qquad (11.5)$$

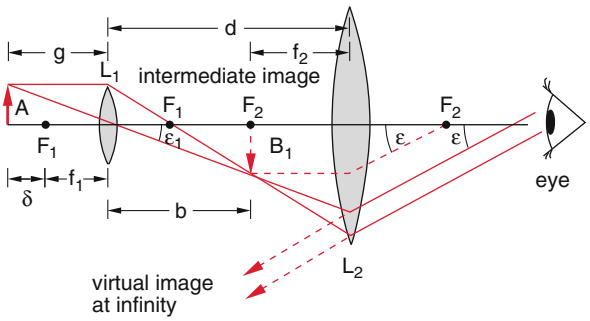


Fig. 11.10 Basic principle of the ray path in a microscope. The intermediate image b appears at the focal plane F_2 of lens L_2

With the distance $d = b + f_2$ between L_1 and L_2 one gets with $g \approx f_1$

$$V_M \approx \frac{(d - f_2)s_0}{f_1 f_2} . \qquad (11.6)$$

Example

$f_1 = 0.5$ cm, $f_2 = 2$ cm, $d = 10$ cm, $s_0 = 25$ cm $\Rightarrow$ $V_M = 200$.

The magnification can be controlled by the choice of the focal lengths f_1 and f_2. Generally different objective lenses are mounted in a rotatable cylinder and can be brought into the light path by rotating the cylinder.

The commercial microscopes are more complicated than the simple principal device in Fig. 11.10. The two single lenses are replaced by lens systems which correct lens aberrations and allow a larger aperture angle. In Fig. 11.11a the ray path is shown for imaging of the object which is placed in front of the microscope objective and which is illuminated by light from a hot tungsten filament. The image appears on the retina of the observing eye. On the right side of Fig. 11.11 the ray path for the illumination of the object is depicted. The bright tungsten filament is imaged into the eye lens of the observer and not onto the retina. The observer therefore does not see the filament but only a bright background at the position of the illuminated object. In Fig. 11.12 the design of a Zeiss microscope is illustrated. The light rays are divided by a beam splitter and two light sources LQ1 and LQ2 are used, which allow the observation of the object with transparent illumination and in reflected light. Instead of the observer's eye a video camera can be installed which can send the pictures directly to a computer [4].

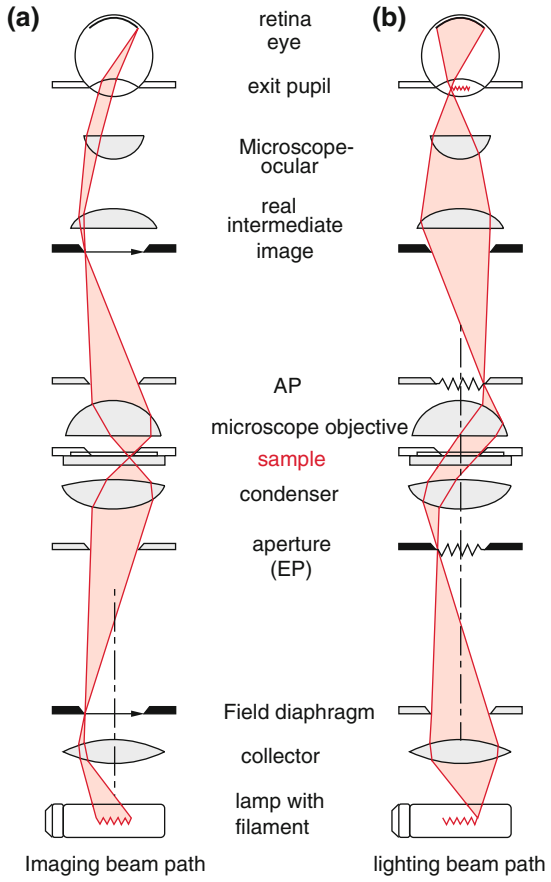


Fig. 11.11 Ray path in the microscope with illumination of the object. **a**) Imaging of the object, **b**) imaging of the illuminating light source (after Pedrotti: Optics.)

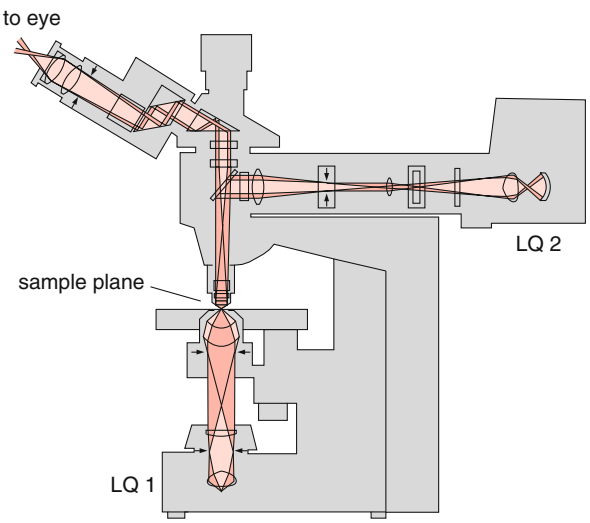


Fig. 11.12 Cut through a commercial microscope (Zeiss, Oberkochen Germany)

11.2.3 Telescopes

Contrary to the microscope which magnifies objects close to the objective lens, telescopes are used to magnify the images of objects at far distances. The first telescope was constructed 1608 in Holland by *Hans Lippershey* (1570–1619) and later Galilei improved it to use ít as astronomical telescope for the observation of the planets (see Vol. 1, Fig. 1.1). Such a telescope was also used in a modified form by Johannes Kepler (1571–1630). The principle of the Kepler telescope is shown in Fig. 11.13. It consist, analog to the microscope, of two lenses. Here, however, the lens L_1 has a large focal length f_1. It generates a real intermediate image of the faraway object in the focal plane of the lens L_1 which coincides with the left

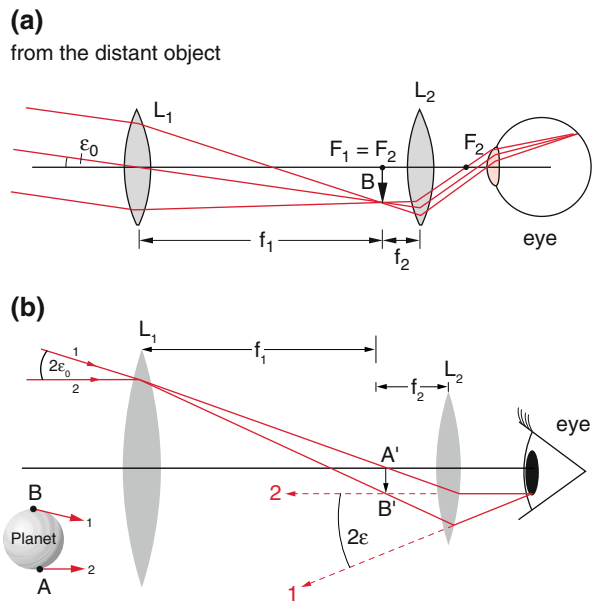


Fig. 11.13 **a**) Magnification by the Kepler- telescope. **b**) Determination of the angular diameter of a planet as angle 2ε between the rays from opposite edges of the planet

Fig. 11.14 Prismatic binocular (with kind permission of Zeiss oberkochen)

focal plane of L_2. The lens L_2 acts as magnifying glass for observing this intermediate image. When the symmetry axis of the telescope (black line in Fig. 11.13) points to the center of a planet with diameter D, the angle between the rays from the edges of the planet is $2\varepsilon_0 = D/f_1$, while the angle between the rays seen by the observer behind L_2 is $2\varepsilon = D/f_2$.

The angular magnification is then

$$V_F = \frac{\varepsilon}{\varepsilon_0} = \frac{B}{f_2\varepsilon_0} = \frac{f_1\varepsilon_0}{f_2\varepsilon_0} = \frac{f_1}{f_2}. \qquad (11.7)$$

The angular magnification of the telescope is equal to the ratio f_1/f_2 of the focal lengths.

Note The image of an object, formed by the Kepler telescope, is inverted.

Example

$f_1 = 2$ m, $f_2 = 2$ cm $\Rightarrow V_T = 100$

Remarks:

(a) Equation (11.6) converts to (11.7) for $d = f_1 + f_2$ and $s_0 = f_1$.

(b) If one wants to avoid the inversion of the image (for example if objects on earth are observed) either inversion prisms can be used (prismatic binocular Fig. 11.14) or the ocular must be a diverging lens (Fig. 11.15).

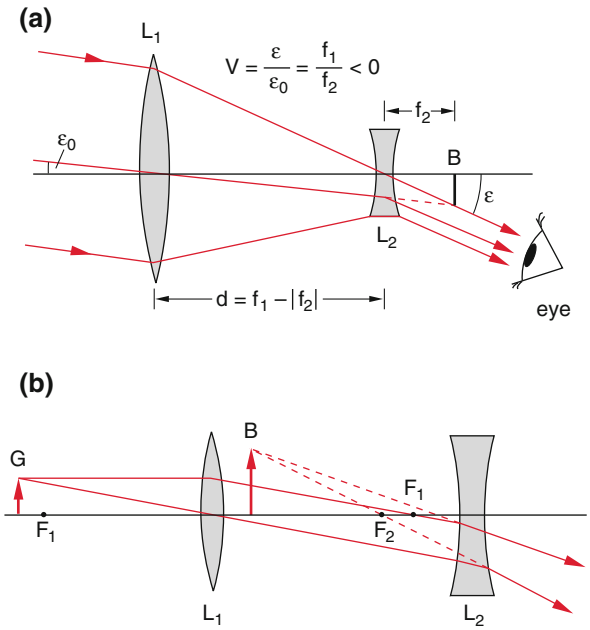


Fig. 11.15 Telescope with diverging lens as ocular. **a**) Angular magnification for an object at infinite distance a = ∞, **b**) generation of an upright image for a finite distance a of the object

For astronomical observations today generally mirror telescopes instead of lens telescopes are used (Fig. 11.16) because spherical or parabolic mirrors can be produced with

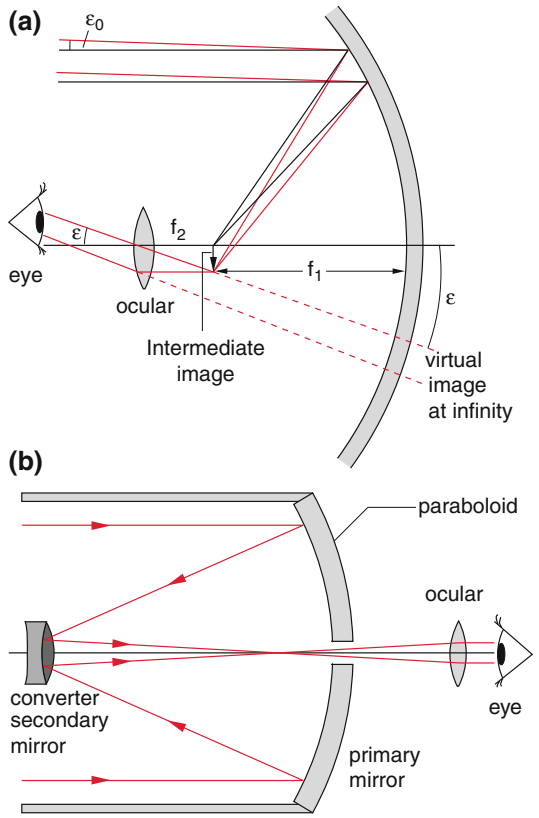


Fig. 11.16 Mirror telescope.**a**) Schematic representation of the angular magnification (in real telescopes the light is deflected into the eye by a small mirror before the ocular, **b**) cassegrain- telescope

much larger diameters than lenses [5]. This increases the luminosity of the telescope which is proportional to the square of the mirror diameter. The largest today existing telescopes (European Southern observatory on the Paranal, a mountain in Chile, and the Keck telescope on Hawaii) have mirror diameters of 10 m. Still larger telescopes are in preparation which consist of many mirrors (up to 200) forming together a parabolic reflecting surface (see Vol. 4, Chap. 10). The alignment of the different mirrors must be precise within $\lambda/10$, which means within 50 nm! This can be only achieved with laser-interferometric techniques [6].

There are several designs of mirror telescopes. In the *Cassegrain* telescope(Fig. 11.16b) the light, falling onto the large primary mirror is reflected onto a small secondary mirror which reflects and focusses the light through a small hole in the primary mirror and an ocular lens images the light into the eye of the observer. Nowadays electronic devices (CCD cameras or cooled CCD arrays) are used as detectors instead of the visual observation.

In the radio frequency range (megahertz to Gigahertz range) huge parabolic antennas (diameters about 100 m) are used for the detection of radio waves from the universe). In Fig. 9.17 the radio telescope in *Effelsberg* illustrates the size of such devices. The parabolic mirror can be turned and tilted to any desired position on the sky within a large angular range.

11.3 The Importance of Diffraction in Optical Instruments

In Sect. 11.2 we have discussed that magnifying optical instruments allow the resolution of finer details of the observed object. The increase of the spatial resolution is, however, limited by diffraction. This will be illustrated by two examples: The telescope and the microscope.

11.3.1 Angular Resolution of Telescopes

We regard in Fig. 11.17b the images of two stars S_1 and S_2 with the angular distance δ generated in the focal plane of a telescope. Because of their large distance the stars can be treated as point light sources. This implies that the light from these stars falling onto the telescope can be regarded as plane waves. Due to diffraction at the limiting aperture D of the telescope the image of a star is no longer a point but a circular intensity distribution $I(r)$ in the image plane, which is shown in Fig. 11.17a as a cut $I(x)$ in the x-direction. The diameter d of the central diffraction maximum is

$$d_{\text{diffr}} = 2f_1 \cdot \sin\alpha_{\text{diffr}} \approx 2.44 \cdot f_1\, \lambda/D. \qquad (11.8a)$$

In Fig. 11.17b the diffraction limited intensity distributions $I(x - x_1)$ and $I(x - x_2)$ of the images of two close stars around the central focal points $F_1(x_1, z_0)$ and $F_2(x_2, z_0)$ are shown in the focal plane $z = z_0$.When the central maximum $I_1(x_1)$ coincides with the first minimum of $I_2(x - x_2)$ the superposition $I(x) = I_1(x) + I_2(x)$ of the two images just barely shows does two distinctly separated maxima, i.e. for smaller separations one cannot decide whether the observed intensity distribution is due to two separated sources S_1 and S_2 (**Rayleigh Criterion**) (see also Problem 11.4 and Sect. 11.5.3). Since the first minimum appears at the diffraction angle $\theta = 1.22\ \lambda/D$ (see Sect. 10.5) the minimum still resolvable angle between the light rays from two sources is

$$\delta_{\min} = 1.22 \cdot \lambda/D. \qquad (11.8b)$$

For this angular distance the superposition of the two Bessel functions $J_1(x - x_1)$ and $J_2(x - x_2)$ has still two distinct maxima at $x = x_1$ and $x = x_2$ with a recess of the total intensity

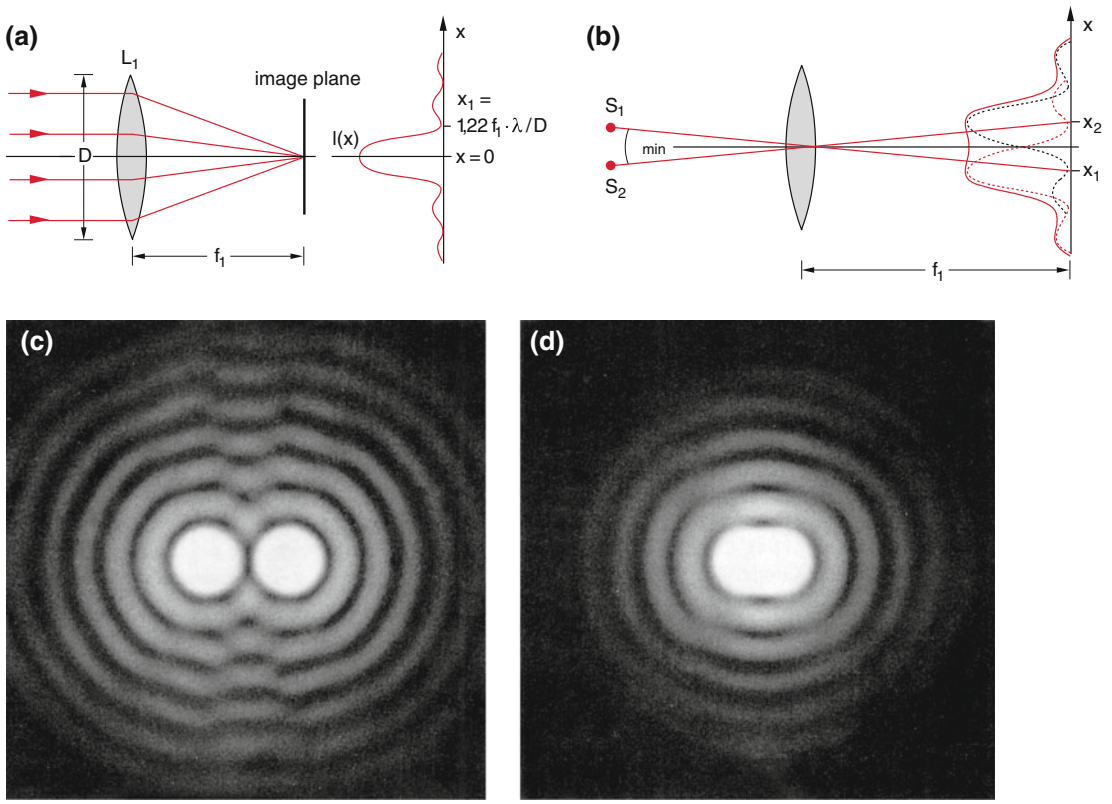


Fig. 11.17 Limitation of the angular resolution of a telescope by the diffraction at the telescope aperture (entrance pupil). **a**) Diffraction intensity distribution in the focal plane of L_1, **b**) superposition of the images of two barely resolvable objects, **c**) photo of two barely resolved point-like sources, **d**) Rayleigh criterion of the diffraction limited resolution

$I(x = ½(x_1 + x_2)) = 0.85 \cdot I_{max}$ between these maxima. We therefore define the quantity

$$R_W = \frac{1}{\delta_{min}} = \frac{D}{1.22\lambda}. \qquad (11.8c)$$

as the **diffraction limited angular resolving power** of the telescope.

The angular resolution of an optical instrument is limited by the ratio D/λ of diameter D of the limiting aperture and wavelength λ

Example

$\lambda = 500$ nm, $D = 1$ m, $f_1 = 10$ m $\Rightarrow \Delta_{min} = 6\ \ 10^{-7}$ rad $= 0.13'' \Rightarrow d_{diffr} = 6$ µm

This diffraction limited resolution is, however, for earthbound telescopes with $D > 10$ cm not the real limiting factor because the random fluctuation of the atmospheric refractive index (air turbulence) limits the angular resolution to about 1″. With a special technique, the **speckle interferometry** [7] or the **adaptive optics** (Sect. 12.3) the air turbulence can be partly outwit, In particular the adaptive optics allows even for large telescopes an angular resolution close to the diffraction limit.

For telescopes outside the earth atmosphere (e.g. the Hubble telescope) air turbulence is of course completely absent and these instruments reach indeed the diffraction limited resolution.

For a mirror diameter $D = 2.4$ m (Hubble telescope) this means for $\lambda = 500$ nm an angular resolution of $\delta_{min} = 2.54\ 10^{-7}$ rad $= 0.052''$. This corresponds to a spatial resolution on the moon (r = 380.000 km) of $\Delta x_{min} = 96$ m!

11.3.2 Resolving Power of the Human Eye

The pupil of the eye has a diameter D which varies according to the incident light intensity between 1 and 8 mm. Neglecting all lens aberrations the eye lens creates on the retina a circular diffraction disc as image of a point like source with a diameter

$$d_{\text{diffr}} = 2.44 f\lambda/D$$

For green light (λ = 550 nm), which converts in the eye ball (refractive index n = 1.33) to λ = 413 nm, this gives for f = 24 mm, D = 2 mm:

$$d_{\text{diffr}} \approx 10\,\mu\text{m}.$$

This corresponds to the mean distance between the photoreceptors (cones and rods) in the area of the fovea in the retina where the packing density of the photo receptors is maximum. This illustrates that nature has optimized in the course of a long development the structure of the retina to match the diffraction limited resolution.

The corresponding diffraction limited angular resolution for λ = 550 nm is

$$\delta_{\min} \approx 1.22\lambda/D \approx 2.9 \cdot 10^{-4}\text{rad} \approx 1'.$$

Our eye can therefore resolve structures of objects at the clear visual range s_0 down to the minimum size of

$$\Delta x_{\min} = s_0 \cdot \delta_{\min} \approx 25\,\text{cm} \cdot 2.9 \cdot 10^{-4} \approx 70\,\mu\text{m}$$

For the resolution of smaller details one needs magnifying glasses or microscopes.

Remark: The resolving power depends also on the form of the object and the contrast of its structure.

11.3.3 Resolving Power of the Microscope

Diffraction represents also for the microscope the principal limit of the spatial resolution.

We regard in Fig. 11.18 a point P_1 of the illuminated object in the observation plane which has the distance g from the object lens L_1 with diameter D.

In the image plane with distance b from L_1 the image of P_1 shows a diffraction pattern with the diameter of the central diffraction maximum

$$d_{\text{diffr}} = 2.44 \cdot \lambda \cdot b/D. \tag{11.9a}$$

In order to recognize a point P_2 with a distance $\Delta x = P_1P_2$ as spatially separated from P_1 the distance between the corresponding diffraction maxima of their images must be at least $\frac{1}{2}d_{\text{diffr}} = 1.22\ \lambda\ b/D$. According to the lens equation this corresponds to a distance between two still resolvable object points

$$\Delta x_{\min} = \frac{1}{2} d_{\text{diffr}} \cdot \frac{g}{b} = 1.22\lambda \cdot \frac{g}{b}. \tag{11.9b}$$

Since the object observed by a microscope lies generally in the focal plane, it is $g \approx f_1$ (Fig. 11.19). This gives for the minimum still resolvable distance between two object points

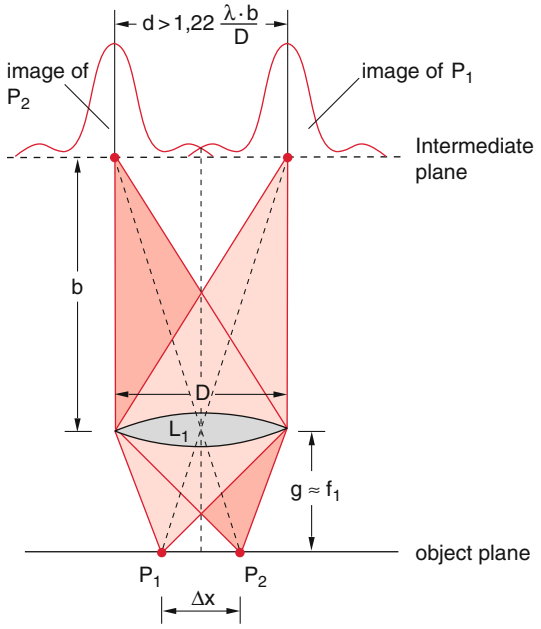


Fig. 11.18 Illustration of the derivation of the resolving power of the microscope

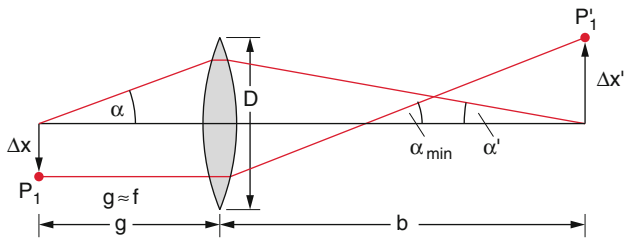


Fig. 11.19 Basic illustration of the characteristic quantities used for the explanation of the diffraction-limited resolution

$$\Delta x_{\min} = 1,22.\frac{\lambda_0}{2n \cdot \sin\alpha}. \tag{11.9c}$$

The image distance is large compared to the object distance ($b >> f_1$). From Fig. 11.18x we obtain the relation

$$\tan\alpha_{\min} = \frac{\Delta x'}{b} \ll 1 \Rightarrow \tan\alpha_{\min} \approx \alpha_{\min}$$

Similarly is

$$\tan\alpha' \approx \alpha' = \frac{D}{2b}$$

$$\Rightarrow \Delta x' \cdot \alpha' = (D/2) \cdot \alpha_{\min} = \frac{D}{2} \cdot \frac{1.22\,\lambda}{D} = 0.61\,\lambda\,. \tag{11.9d}$$

The image equation follows Abbe's sin-theorem (9.40)

$$\Delta x \cdot \sin\alpha = x' \cdot \sin\alpha' \approx \Delta x' \cdot \alpha' \tag{11.9e}$$

The maximum aperture angle α collected by the objective lens L_1 is given by

$$\Delta x \cdot \sin\alpha = \Delta x' \cdot \sin\alpha' \approx \Delta x' \cdot \alpha' \qquad (11.10)$$

We can then write (11.9c) as

$$\Rightarrow \Delta x_{\min} = \frac{\Delta x' \cdot \alpha'}{\sin\alpha} = \frac{0.61\lambda}{\sin\alpha}. \qquad (11.9f)$$

Using immersion oil with a large refractive index (n = 1.5) between object and objective lens the resolution can be increased by the factor 1.5 because $\lambda_n = \lambda_0/n$ This gives for the minimum distance

$$\Delta x_{\min} = 1.22 \cdot \frac{\lambda_0}{(2n \cdot \sin\alpha)}. \qquad (11.11a)$$

The product $n\ sin\alpha = NA$ is called the **numerical aperture** of the microscope . We can write (11.11a) then as

$$\Delta x_{\min} = 0.61 \frac{\lambda_0}{NA}. \qquad (11.11b)$$

Example

n = 1.5, $\sin\alpha$ = 0.8, $(2\alpha = 106°) \Rightarrow NA = 1.2 \Rightarrow x_{min} \approx 0.5\ \lambda$.

Structures on objects illuminated by light with the wavelength λ, which are smaller than λ/2 cannot be resolved.

In order to reach a higher spatial resolution for optical microscopes the wavelength λ has to be decreased. Meanwhile first successful attempts for the construction of X-ray- microscopes ($\lambda \approx$ 10–50 nm) with Fresnel lenses (see Sect. 12.6.2) have been performed. New optical techniques with lasers can overcome the diffraction limit (see Sect. 11.3.5).

With electron microscopes (see Vol. 3) a much higher spatial resolution down to Δx = 0.1 nm) can be realized.

11.3.4 Abbe's Theorem of the Formation of Images

Ernst Abbe (1840–1905) realized already 1890 that diffraction plays an essential role for the spatial resolution of imaging optical instruments. He illustrated his theory by the example of image formation in a microscope (Fig. 11.20).

We regard two illuminated slits S_1 and S_2 with the distance d as our objects. If only the zeroth diffraction order is observed no information is obtained about the distance d.

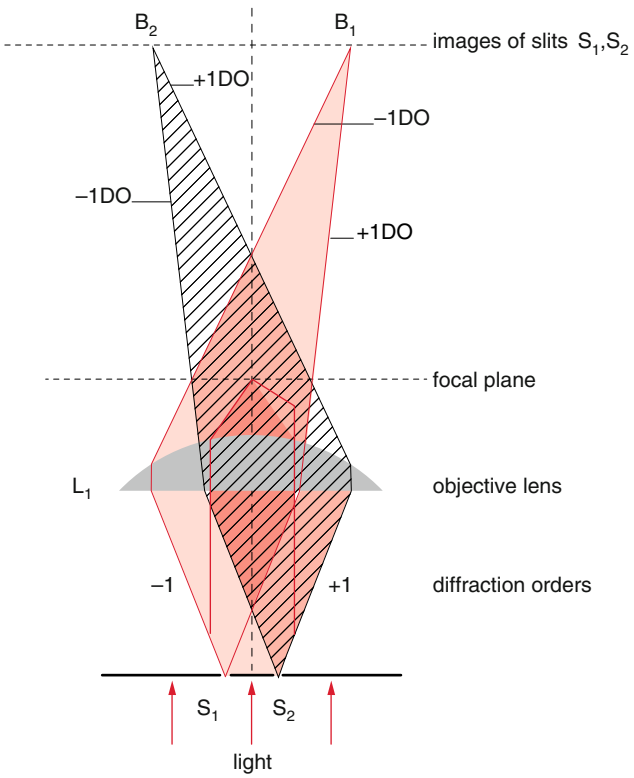


Fig. 11.20 Illustration of Abbe's theory of image formation in a microscope. The red area gives the transmitted light of S_1 between the +1. and the −1. diffraction orders, the black shaded area that of light source S_2

The mth diffraction order appears under the angle θ_m against the propagation direction of the incident light. Since

$$d \cdot \sin\theta_m = m \cdot \lambda \quad (m = 1, 2, 3, \ldots)$$

the diffraction angle $\theta_m = \arcsin(m\ \lambda/d)$ depends on the distance d. Figure 11.20 illustrates that at least the +1. and the −1. diffraction order contribute to the generation of the images of the object slits. The objective lens therefore must collect at least these diffraction orders to form the images of the objects. This implies that the numerical aperture NA must be at least

$$NA = n \sin\alpha > n \sin\theta_1 = \frac{\lambda}{d} \qquad (11.12a)$$

(where n is the refractive index of the immersion oil) in order to achieve the spatial resolution $\Delta x_{min} = d$. The minimum distance of two still resolvable objects is then for a given numerical aperture

$$d_{\min} = \frac{\lambda}{(n \sin\alpha)} = \frac{\lambda}{NA}, \qquad (11.12b)$$

This agrees, apart from a factor 0.6 with (11.11b).

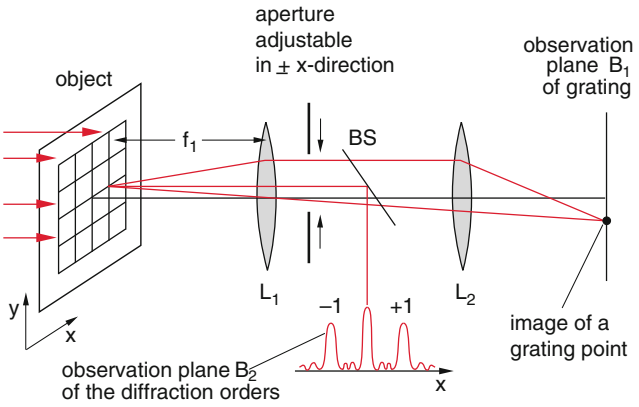


Fig. 11.21 Demonstration of Abbe's theory

The Abbe theory can be impressively demonstrated by the following experiment: A quadratic grid in the *x*-*y* focal plane of L_1 is illuminated by parallel light (Fig. 11.21). Behind L_1 two mutual perpendicular slits with variable slit width are placed in the *x*-direction resp. the *y*- direction. If one of the slits is constricted such, that only the zeroth diffraction order of the illuminated grid is transmitted, the grid structure in one direction disappears and the image of the two-dimensional quadratic grid becomes a one-dimensional graticule, where the lines in the image are perpendicular to the narrowed slit. If also the second slit is narrowed the grid structure of the image disappears completely. With a beam splitter BS part of the light can be reflected onto the plane B_2 where the *Fraunhofer* diffraction structure of the grid can be observed and one can see, which diffraction orders are transmitted by the slits.

Within the framework of Fourier-representation (Sect. 10.8) the Fraunhofer diffraction intensity distribution can be regarded as Fourier-transform of the field distribution in the diffraction plane. The image of the object in the observation plane B_1 is the Fourier transform of the diffracted intensity distribution. If spatial structures are missing (because they have been cut off by the aperture) the corresponding Fourier parts are missing in the real image, which means that the image is washed out (see Chap. 12).

11.3.5 Surpassing of the Classical Diffraction Limit

In the last years several methods have been developed, which can surpass the resolution limits, previously regarded as a fundamental bound. We will here discuss only some of them:

(a) Confocal microscopy (Sect. 12.1)
(b) Optical near field microscopy (Sect. 12.2)
(c) the 4π-microscopy (this section)
(d) Stimulated depletion spectroscopy (this section).

All these techniques do not contradict Abbe's theory, because they achieve a higher spatial resolution by using some tricks:

They either use a spatial narrowing of the light emitted by the illuminated object, by using spatial filters (a), or they limit the spatial volume of the light emitting molecules (b), or they use the destructive interference when a laser beam superimposes the beam reflected by a spherical mirror (c), or they suppress the light emitted by the illuminated object for all locations except a very small volume in the center of the illuminating laser beam (d), which increases the spatial resolution at least by a factor 10.

All these techniques use the high intensity and the coherence properties of laser light sources (see Vol. 3, Chap. 8). They therefore could not have been developed at Abbe's time.

In many cases not only the lateral spatial resolution (perpendicular to the axis of the microscope) but also the axial resolution (in the direction of light propagation) play the essential role. For the classical microscope this is the Rayleigh length z_R ($2z_R$ is the distance around the focus, where the diameter of the focused light beam has increased to $\sqrt{2}$ times the diameter $2w_0$ in the focal plane (Fig. 11.22). The cross section of the light beam has then doubled compared to $\pi\, w_0^2$ in the focal plane. As can be shown [8]:

The Rayleigh length is $z_R = \pi \cdot w_0^2/\lambda$.

Inserting for w_0 the minimum still resolvable distance $\Delta x = 0.7\ \lambda$ in air ($n = 1$) we obtain a Rayleigh length $z_R = 1.5\lambda$.

The technique of 4π microscopy, developed by Stefan Hell [9] improves the axial resolution considerably.

The coherent laser light is focused by the lens L1 into the sample. It is then reflected back into itself by a spherical mirror and again focused into the same spot as the incident light (Fig. 11.23). The superposition of the two waves gives the total field amplitude

$$E(z, r, \phi) = E_1(z, r, \phi) + E_2(-z, r, \phi) \qquad (11.13a)$$

of a standing wave with the intensity distribution

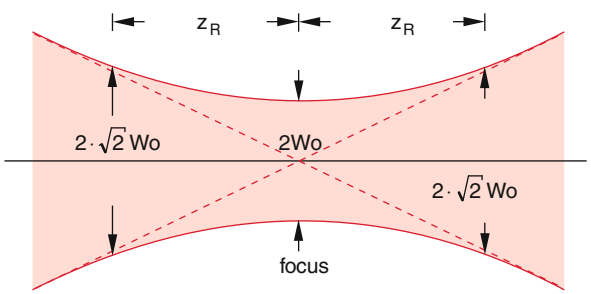


Fig. 11.22 Definition of the Rayleigh length

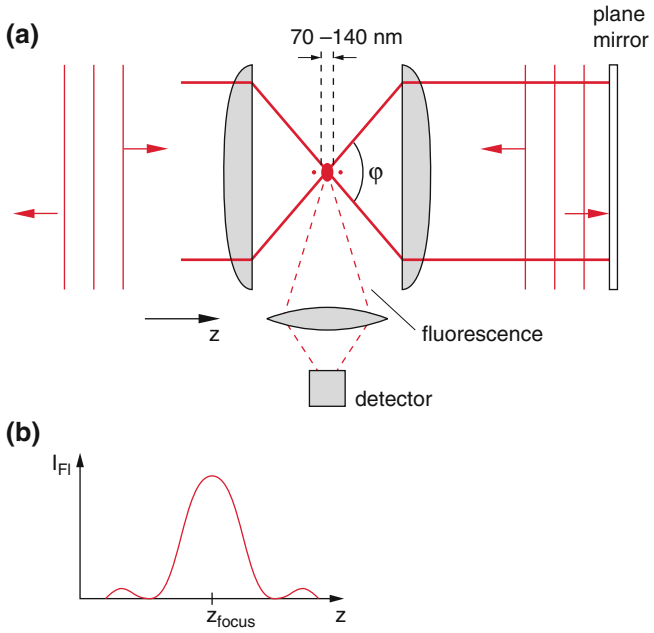


Fig. 11.23 **a**) The 4π imaging method to improve the axial resolution, **b**) intensity $I(z)$ of the fluorescence emitted by the sample in the focal region

$$I(z, r, \phi) = |E_1 + E_2|^2 = E_1^2 + E_2^2 + 2E_1 \cdot E_2 \cdot \cos(kz + \phi). \quad (11.13b)$$

The last term describes the interference between incident and reflected beam, which leads to intensity maxima and minima with the distance of $\lambda/2$. Because of the strong divergence of the focused beam the intensity decreases strongly as $1/r^2$ with increasing distance z from the focus. The intensity $I(z)$ has a narrow maximum at the focus z = 0 with the half width $\Delta z \approx \lambda/4$. For the next maximum at $z = \lambda/2$ the beam radius has already decreased from w_0 in the focus to

$$w(z = \lambda/2) = w_0 \cdot \left[\left(\frac{1 + 2f^2}{\pi w_1^2}\right)^2\right]^{1/2} \quad (11.13c)$$

where f is the focal length of the lens L_1 and w_1 is the beam diameter at the lens. For typical values $f = 2\ w_1$ one obtains r $(z = ½\ \lambda) = 3.6\ w_0$. The intensity $I(z = ½\ \lambda)$ of the next maximum has already decreased to 0.077 of the intensity in the first maximum at z = 0.

Often a two-photon absorption spectroscopy is used. Here the intensity decreases with $1/r^4$ and the second maximum has only 0.6% of the intensity in the central maximum.

The signal is the fluorescence emitted from the excited molecules in the focal volume. It is observed through a microscope perpendicular to the incident laser beam.

With this technique it is possible to resolve structure in biological cells that is smaller than $\lambda/10$.

The radial resolution can be greatly improved be the technique of stimulated depletion spectroscopy [10]. Here the molecules are excited in the focus of a Gaussian laser beam and emit fluorescence. The excited molecules are de-excited by a second "depletion laser" with a radial donut beam profile that surrounds the Gaussian profile of the first laser. This depletion laser induces stimulated downward transitions of the excited molecules and extinguishes the fluorescence because the stimulated emission propagates into the direction of the laser beam and does not reach the detector, perpendicular to the laser beam. Depending on the intensity of the depletion laser the radial profile of the fluorescence emitting molecules can be narrowed down to 10 nm.

Another method for the improvement of the radial and lateral resolution is the confocal microscopy which is discussed in Sect. 12.2 (see also [11]).

11.4 The Luminosity of Optical Instruments

Besides their spatial resolution the luminosity of optical instruments is often essential for many applications. Examples are cameras, telescopes, slide projectors spectrographs etc.

The light power transmitted by an optical instrument depends on the apertures which limit the cross section of the transmitted light beam. Such apertures can be the lens mounts, the area of prisms or gratings in spectrographs or additional apertures in the light path through the instrument.

We name the common cross section of all incident light rays the **entrance pupil**, the common cross section on the image side the **exit pupil**.

For the simple imaging of an extended object by the lens L (Fig. 11.24) the lens cross section is the common entrance- and exit pupil. Placing an aperture B in the object space before the lens (Fig. 11.25), this aperture limits the maximum opening angle Ω for light emitted by any point P of the object and therefore also limits the transmitted light power. The aperture acts as entrance pupil. The real image of the aperture on the image side is the exit pupil which

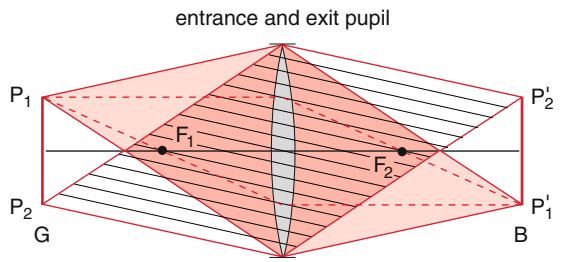


Fig. 11.24 For the imaging by a lens without further apertures the lens mount is the common entrance and exit pupil

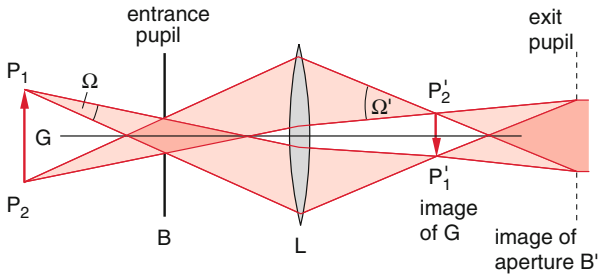


Fig. 11.25 The aperture B on the object side represents the entrance pupil, its image on the image side acts as exit pupil

transmits only light beams on the image side with opening angles $\leq \Omega'$.

Since a self-luminous object generally emits light into the whole solid angle 4π, the light power transmitted by the instrument is proportional to the opening angle Ω accepted by the entrance pupil. If the object is placed in the focal plane of the light collecting lens with the focal length f and the entrance pupil with diameter D in the plane of the lens L (e.g. for a camera) the accepted solid angle is for $D < f$

$$\Omega = \frac{\pi D^2/4}{f^2} = \frac{\pi}{4}\left(\frac{D}{f}\right)^2. \qquad (11.13)$$

Enlarging the diameter D by the factor $\sqrt{2}$ increases the transmitted light power by the factor 2. For cameras the label "aperture 8" means $f/D = 8$. The ratio f/D is often called "*f-number* ". For $f/D = 8$ and $f = 40\,\text{mm} \Rightarrow D = 5\,\text{mm}$, for the f-number 11 is $D = 3.6\,\text{mm}$ and the transmitted light power is just ½ of that transmitted for the f-number 8.

We will illustrate the luminosity for the example of a projector (Fig. 11.26). The light emitted by a bright lamp is collected by the condenser lens with a preferably large ratio D/f. In order to collect also the light emitted into the backwards direction a spherical mirror reflects the light back into the lamp and to the condenser lens. The slide is placed at a location where the light beam has about the same cross section as the slide in order to uniformly illuminate the slide. Each point of the slide is now imaged by the lens L_2 (objective, often a system of several lenses) onto the projection screen. For a size G of the slide the size of the projection image is

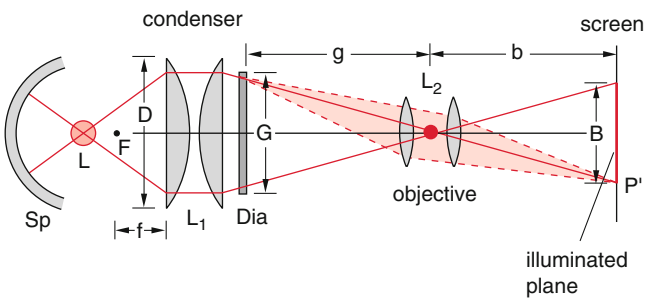


Fig. 11.26 Optical design of a slide projector

$$B = \frac{b}{g}G \qquad (11.14b)$$

where b is the distance between projection screen and the principal plane on the image side of the lens system L_2 and g is the distance between slide and principal plane of L_2 on the object side. The two quantities b and g are not independent but are related by the lens equation

$$\frac{1}{f_2} = \frac{1}{g} + \frac{1}{b} \qquad (11.14c)$$

The image of the bright filament of the lamp should not be imaged onto the screen. This can be avoided by the proper choice of f_1 and f_2. Generally the focal lengths are chosen such,that the image of the filament lies between the lenses of the lens system L_2 (red point in Fig. 11.26).

11.5 Spectrographs and Monochromators

The spectral distribution $I(\lambda)$ of the light emitted by a radiation source can be measured with spectrographs where the transmitted light is spatially separated according to its wavelength λ [12]. This spatial separation can be achieved with prisms (**prism spectrograph** Fig. 11.27) or with optical gratings (**grating spectrograph**, Fig. 11.28).

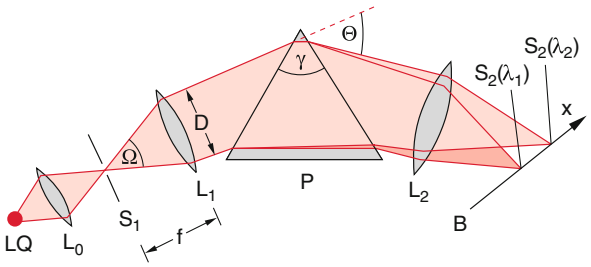


Fig. 11.27 Prism-spectrograph

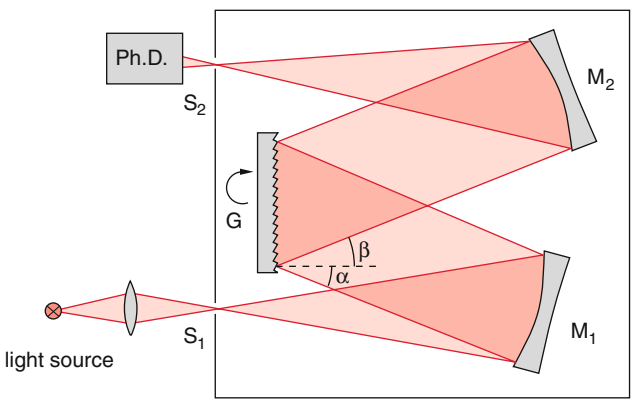


Fig. 11.28 Grating monochromator(PhD = Photodetector)

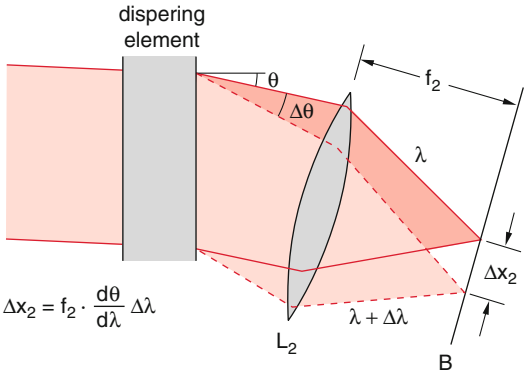


Fig. 11.29 Relation between angular and spectral dispersion

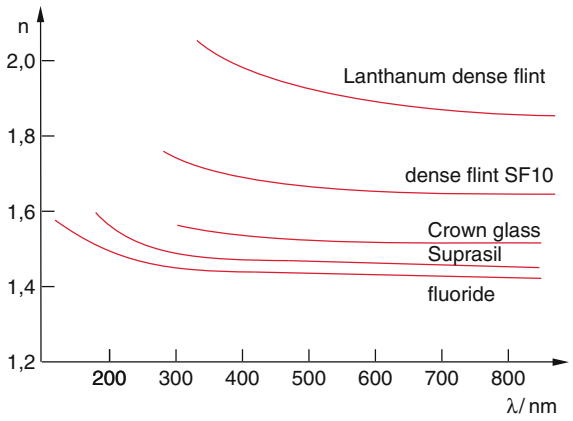


Fig. 11.30 Dispersion curves $n(\lambda)$ for some optical materials

In prism spectrographs the dispersion of the refracted index of the prism material is used which leads to wavelength-dependent refraction angles, while in grating spectrographs the wavelength-dependent diffraction and interference of the waves reflected by the optical grating cause the spatial separation of the reflected light.

In all types of spectrographs an entrance slit S_1 is imaged by lenses or mirrors onto the observation plane. If a slit S_2 with width Δx is placed in the observation plane it transmits only a limited wavelength interval

$$\Delta\lambda = (\mathrm{d}\lambda/\mathrm{d}x)\Delta x$$

which is determined by the inverse wavelength dispersion $(\mathrm{d}x/\mathrm{d}\lambda)^{-1} = \mathrm{d}\lambda/\mathrm{d}x$ of the spectrograph. The spectrograph has been converted into a monochromator. The wanted wavelength can be transmitted either by shifting the slit across the observation plane or by turning the grating in Fig. 11.28 around a vertical axis.

The dispersing element (prism or grating) causes a wavelength-dependent deflection θ of the incident parallel light beam (Fig. 11.29). The lens L_2 (or the mirror M_2 in the grating spectrograph) focuses the parallel light into the observation plane. The lateral shift $x(\lambda)$ of the slit image $S_2(\lambda)$ is for the wavelength change $\Delta\lambda$

$$\Delta x = x(\lambda + \Delta\lambda) - x(\lambda) = f_2 \frac{\mathrm{d}\theta}{\mathrm{d}\lambda}\Delta\lambda \qquad (11.14\mathrm{d})$$

It depends on the angular dispersion $\mathrm{d}\theta/\mathrm{d}\lambda$ and the focal length f_2 of the objective lens L_2.

11.5.1 Prism Spectrographs

The light emitted by the light source LQ in Fig. 11.27 is collected by the lens L_0 and focused onto the entrance slit S_1which is placed in the focal plane of the lens L_1. The light emerging through the slit S_1 is formed by L_1 into a parallel light beam which passes through the prism. Due to the wavelength-dependent dispersion the different wavelengths suffer a different diffraction and form behind the prism different parallel light beam with different deflection angles θ. The lens L_2 produces in the observation plane spatially separated slit images $S_2(\lambda_i)$ for the different wavelength λ_i. The deflection angle $\theta(\lambda)$ is for the symmetric optical path through the prism according to (9.20) given by

$$\frac{\mathrm{d}\theta}{\mathrm{d}\lambda} = \frac{2\sin(\gamma/2)}{\sqrt{1 - n^2\sin^2(\gamma/2)}} \cdot \frac{\mathrm{d}n}{\mathrm{d}\lambda} \qquad (11.14\mathrm{e})$$

It depends on the prism angle γ and the dispersion $\mathrm{d}n/\mathrm{d}\lambda$ of the prism material (Fig. 11.30).

Example

For flint glass is for $\lambda = 500\,\mathrm{nm}$ $n = 1.81$, $\mathrm{d}n/\mathrm{d}\lambda = 4400/\mathrm{cm}$. With an equilateral prism this gives: $\mathrm{d}\theta/\mathrm{d}\lambda = 1.02 \cdot 10^{-3}\,\mathrm{rad/nm}$. The slit images for two wavelengths that differ by $\Delta\lambda = 10$ nm are separated by $\Delta x = f_2\,\mathrm{d}\theta/\mathrm{d}\lambda\,\Delta\lambda$. With $f_2 = 40\,\mathrm{cm}$ this gives $\Delta x = 4.1$ mm.

The advantage of the prism spectrograph is its compact design and the unambiguous assignment of the wavelength λ_i from its position $x(\lambda_i)$ in the observation plane. Its drawback is the relative small wavelength dispersion, causing a moderate spectral resolution.

Example

With a slit width $b = 100\,\mu\mathrm{m}$ of the entrance slit two spectral lines can be barely separated in a prism spectrograph with the data of the previous example, if their wavelength difference $\Delta\lambda$ is at least $\Delta\lambda \geq f_2 \cdot \mathrm{d}\theta/\mathrm{d}\lambda \cdot b = 0.25\,\mathrm{nm}$. This gives a relative spectral resolution $\lambda/\Delta\lambda = 500/0.25 = 2000$

Table 11.1 Transmission range of some optical glasses

Glass	Transmission range (nm)
Fused quartzglass	200–3000
Borosilcate glass	350–2000
Crownglass	350–2000
Flintglass	400–2500
dense Flint SF6	380–2500

The measured spectral ranges are restricted to regions where the prism material does not absorb. In the ultraviolet range fused quartz (suprasil) is used. For the infrared region LiF or NaCl are good candidates. Since the dispersion $dn/d\lambda$ is particularly large close to absorbing transitions, one must make a compromise between high spectral resolution and high transmission. In Table 11.1 some materials which are generally used, are listed.

11.5.2 Grating Monochromator

In a grating monochromator (Fig. 11.28) divergent light transmitted through the entrance slit S_1 is collected by the spherical mirror M_1 which reflects a parallel light beam, if $S_1$is located in the focal plane of M_1. The parallel light beam impinges onto the optical reflection grating under the angle α against the grating normal (Fig. 10.45). The different partial waves reflected by the different grating grooves superimpose and interfere constructively in those directions β which fulfill the grating equation

$$d(\sin\alpha + \sin\beta) = m \cdot \lambda \qquad (11.15)$$

For a fixed incidence angle α the direction β for constructive interference depends on the wavelength λ. The reflected parallel light beam is focused by the spherical mirror M_2 onto the exit slit S_2. The photodetector PhD sits behind S_2.

The angular dispersion $d\beta/d\lambda = (d\lambda/d\beta)^{-1}$ is obtained by differentiating (11.15) with respect to β. This gives

$$\frac{d\beta}{d\lambda} = \frac{m}{d.\cos\beta} = \left(\frac{d^2\cos^2\alpha}{m^2} + \frac{2d\lambda}{m}\sin\alpha - \lambda^2\right)^{-1/2}. \qquad (11.16)$$

This illustrates that the angular dispersion depends on the grating constant d (distance between adjacent grooves), the interference order m, the wavelength λ and the angle of incidence α.

The spatial separation of two wavelengths λ_1 and $\lambda_2 = \lambda_1 + \Delta\lambda$ in the observation plane is

$$\Delta x = f_2.\frac{d\beta}{d\lambda}\Delta\lambda = f_2\frac{m.\Delta\lambda}{d\cdot\cos\beta}. \qquad (11.17)$$

11.5.3 The Spectral Resolution of Spectrographs

The spectral resolution is defined as the ratio $\lambda/\Delta\lambda$ of wavelength λ and the minimum still resolvable wavelength interval $\Delta\lambda$. For spectrographs is $\Delta\lambda = \lambda_1 - \lambda_2$ the minimum distance between the two wavelengths λ_1 and λ_2 for which two separated slit images can be obtained. Without diffraction the image of the entrance slit with width b is a rectangular intensity distribution with width $B = (f_1/f_1)\ b$ (Fig. 11.31a), where f_1 and f_2 are the focal lengths of the lenses L_1 and L_2 in Fig. 11.27 resp. the mirrors M_1 and M_2 in Fig. 11.28. For most spectrographs is $f_1 = f_2 \Rightarrow B = b$.

Due to the diffraction at the entrance pupil with diameter a (this can be the lens mount of L_1 In Fig. 11.32 or the edges of the mirror M_1 in Fig. 11.28) the intensity distribution $I(x)$ of the entrance slit image in the observation plane is, even for $b \to 0$, not a delta function but becomes the diffraction pattern of Eq. (10.43) in Fig. 11.31c with a bottom width $\Delta x_B = 2 \cdot f_2 \cdot \lambda/a$.

The images of the entrance slit for two adjacent wavelengths λ_1 and $\lambda_2 = \lambda_1 + \Delta\lambda$ are the two distributions $I_1(x_1, \lambda_1)$ and $I_2(x_2, \lambda_2)$ which are separated by the distance $\Delta x = x_2 - x_1$. They can be recognized as separated structures if the maximum of $I_1(x)$ coincides with the first minimum of $I_2(x)$. In this case the superposition of the two distributions shows a recess between the two maxima which is $8/\pi^2$ 0.8 = 80% of the maxima (Fig. 11.33).

Note: Although the diffraction at the much small entrance slit with width b is much larger than that at the entrance

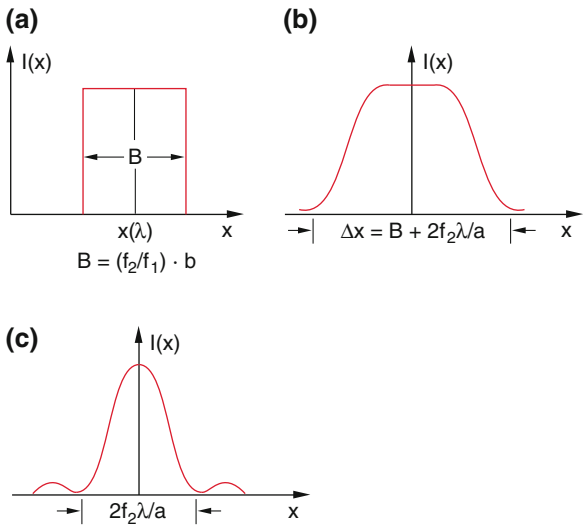


Fig. 11.31 Intensity profile $I(x)$ **a**) without diffraction for finite slit width $b > > \lambda$. **b**) With diffraction for $b \to 0$

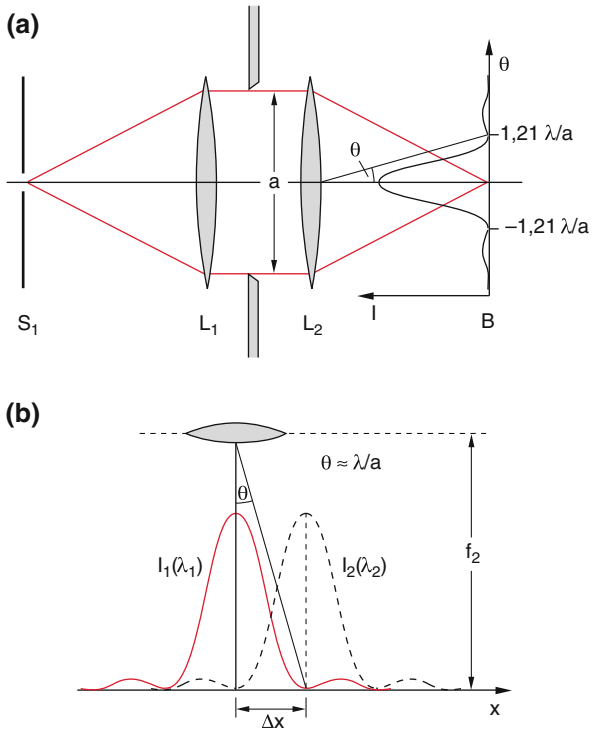


Fig. 11.32 Broadening of slit image by diffraction at the limiting edges of the enlarged light beam

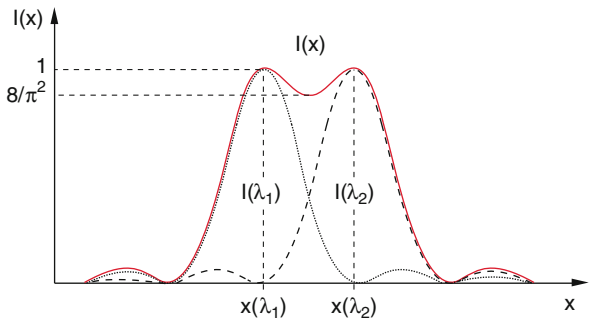


Fig. 11.33 Rayleigh criterion of the resolution of two close spectral lines $I(\lambda_1)$ and $I(\lambda_2)$

pupil with diameter $a \gg b$, it does not limit the resolution. It causes a larger opening angle (in addition to the geometrical divergence). For a parallel light beam falling onto the entrance slit, the light transmitted by the slit has the angular distribution shown in Fig. 11.34 with a diffraction angle $\theta = \lambda/b$ for the half angular width of the central diffraction maximum. If θ becomes larger than the acceptance angle $\alpha/2 = a/2f_1$ of the spectrometer the collimator lens L_1 cannot collect all the light, i.e. the transmitted light power decreases. This is the case for

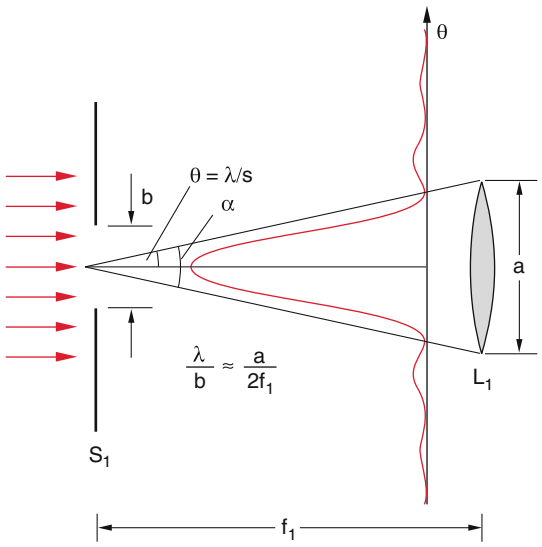


Fig. 11.34 The diffraction by the entrance lit results in a loss of transmission for $\lambda/b > a/(2f_1)$ but not in a broadening of the slit image

$$b < 2f_1 \cdot \lambda/a \tag{11.17a}$$

Therefore b should be larger than $2f_1/\lambda/a$ in order to avoid a noticeable transmission loss. For $b = 2f_1 \cdot \lambda/a$ the slit image in the observation plane, broadened by the diffraction at the entrance pupil has the base width

$$\Delta x = (f_1 + f_2)\frac{\lambda}{a}. \tag{11.18a}$$

With increasing width b of the entrance slit the geometrical image of the slit image in the observation plane B becomes larger. The half-bottom width of the central diffraction maximum becomes for monochromatic light and $f_1 = f_2 = f$ (Fig. 11.31b)

$$\Delta x = \frac{b}{2} + f\frac{\lambda}{a}. \tag{11.18b}$$

This corresponds to a wavelength interval

$$\Delta\lambda = \frac{\mathrm{d}\lambda}{\mathrm{d}x}\Delta x = \left(\frac{1}{f}\right)\frac{\mathrm{d}\lambda}{\mathrm{d}\theta}\Delta x. \tag{11.18c}$$

For the minimum slit width $b = 2f \cdot \lambda/a$ the spectral resolution becomes

$$\frac{\lambda}{\Delta\lambda} = \frac{a}{2}\frac{\mathrm{d}\theta}{\mathrm{d}\lambda}, \tag{11.19}$$

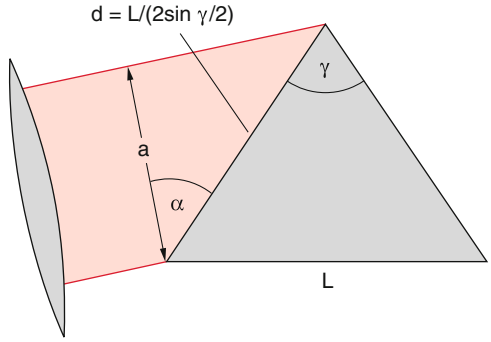


Fig. 11.35 Determination of the entrance pupil with diameter d in the prism spectrograph if the prism represents the boundary of the light beam

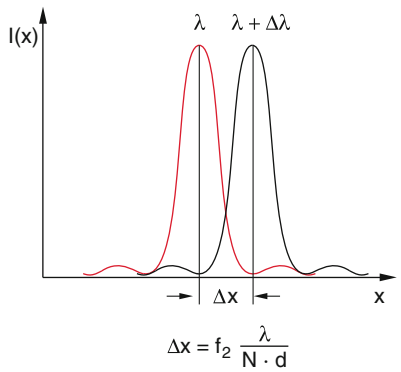


Fig. 11.36 Spectral resolution of the grating spectrograph

For a prism spectrograph with an equilateral prism ($\gamma = 60° \Rightarrow \sin(\gamma/2) = 1/2$) and symmetric optical path we get

$$\frac{\lambda}{\Delta\lambda} = \frac{a}{2}\frac{\mathrm{d}n/\mathrm{d}\lambda}{\sqrt{1-n^2/4}} \tag{11.20}$$

If the entrance pupil is limited by the size of the prism the diameter a of the entrance pupil with base length L is $a = L\ \cos\alpha/(2\sin\gamma/2) = L_2$ for $\alpha = 60°$ (Fig. 11.35). The exit pupil has the same size for symmetric optical path. The spectral resolution of the prism spectrograph Fig. 11.36

$$\frac{\lambda}{\Delta\lambda} = \frac{1}{4}\frac{L}{\sqrt{1-n^2/4}}\frac{\mathrm{d}n}{\mathrm{d}\lambda} \tag{11.21}$$

is then limited by the size L of the prism and the dispersion $\mathrm{d}n/\mathrm{d}\lambda$ of the prism material.

Example

$L = 10\,\mathrm{cm}$; $n = 1.47$ (quartz glass suprasil) $\mathrm{d}n/\mathrm{d}\lambda = 1100/\mathrm{cm} \Rightarrow$

$$\frac{\lambda}{\Delta\lambda} = \frac{1}{4}\frac{10}{\sqrt{1-0,54}} \cdot 1100 = 4060. \tag{11.21a}$$

This means: For $\lambda = 540$ nm two wavelengths can be separated if their difference is at least $\Delta\lambda = 0.14$ nm.

A much higher spectral resolution can be achieved with grating spectrographs. Here the diameter of the exit pupil is $a = N \cdot d \cdot \cos\beta$, where d is the distance between adjacent grooves and N is the number of illuminated grooves (Figs. 11.28 and 11.37a). The angular distance $\Delta\beta$ between the directions of the two wavelengths λ_1 and $\lambda_2 = \lambda_1 + \Delta\lambda$ in the reflected light must be larger than half of the bottom width of the central diffraction maximum for the diffraction at the effective grating width $N \cdot d \cdot \cos\beta$, i.e.

$$\Delta\beta_{\min} = \frac{\lambda}{a} = \frac{\lambda}{N \cdot d \cdot \cos\beta} \tag{11.22}$$

With (11.16) we get

$$\begin{aligned} \Delta\lambda &= \frac{\mathrm{d}\lambda}{\mathrm{d}\beta}\Delta\beta = \frac{d\cos\beta}{m} \cdot \Delta\beta \\ &\geq \frac{d \cdot \cos\beta}{m} \cdot \Delta\beta_{\min} = \frac{\lambda}{m \cdot N} \\ &\Rightarrow \frac{\lambda}{\Delta\lambda} \leq m \cdot N \end{aligned} \tag{11.23}$$

The spectral resolving power of a grating spectrograph with N illuminated grooves is equal to the product of interference order m and the number N of illuminated grooves.

Two wavelengths λ_1 and $\lambda_2 = \lambda_1 + \Delta\lambda$ can be still resolved (for an entrance slit width $b \to 0$) if the maxima of the intensity distributions of the slit images in the observation plane (Fig. 11.35) are separated by at least

$$\Delta x \geq f_2 \cdot \cos\beta \cdot \Delta\beta_{\min} = f_2\lambda/(N \cdot d) \tag{11.23a}$$

Example

A grating with 10 cm width and 1200 grooves/mm operated in second order (m = 2) has a spectral resolution (for full illumination of the grating)

$\lambda/\Delta\lambda = 2 \cdot 1.2 \cdot 10^5 = 2.4 \cdot 10^5$. This is 50 times as large as for the example of the prism spectrograph.

11.5.4 A General Expression for the Spectral Resolution

The Rayleigh criterion for the spatial separation of the slit images of two spectral lines (the maximum of the intensity distribution of the slit image for the wavelength λ_1 should be not closer to that of λ_2 than the first diffraction minimum for λ_2) can be formulated in a more general way:

A maximum of $I(\lambda_1)$ occurs if the maximum path difference Δs_m between the interfering partial waves is an integer multiple of the wavelength λ_1.

$$\Delta s_m = 2q\lambda_1 (q = integer). \tag{11.24a}$$

For this case the total light beam can be divided into two halves.For each partial beam in the first half there exists a partial beam in the second half with a path difference $q \cdot \lambda$, i.e. all partial waves interfere constructively. For the grating spectrograph used in first order (m = 1) is for example $2q = N$.

If for the second wavelength λ_2 the first diffraction minimum should occur at the same diffraction angle ß the condition for destructive interference is

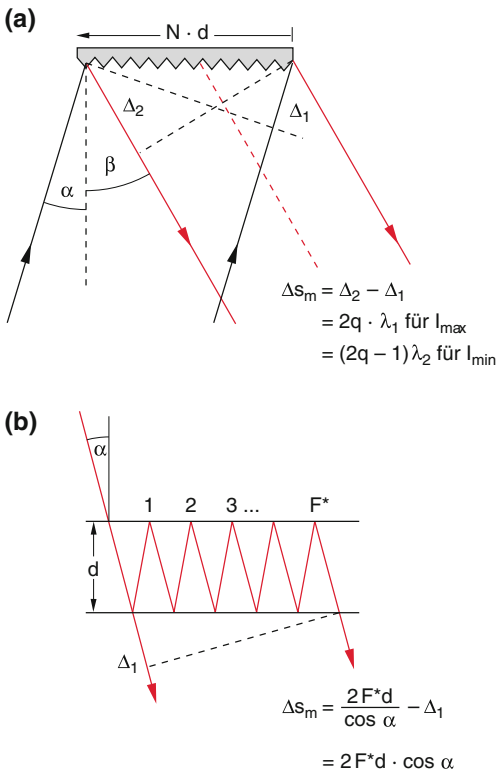


Fig. 11.37 General definition of the spectral resolution $\lambda/\Delta\lambda = \Delta s_m/\lambda$. **a**) For the grating spectrograph is $\Delta s_m = N \cdot m \cdot \lambda$, **b**) for the interferometer is $\Delta s_m = 2F^* \cdot d \cdot \cos\alpha$, where F^* is the finesse

$$\Delta s_m = (2q - 1)\lambda_2. \tag{11.24b}$$

with $\lambda = \sqrt{(\lambda_1 \cdot \lambda_2)}$ we obtain from (11.24a, 11.24b)

$$\frac{\lambda}{\Delta\lambda} = \frac{\Delta s_m}{\lambda}. \tag{11.25a}$$

The spectral resolution is equal to the maximum path difference Δs_m measured in units of the wavelength λ.

With $\lambda = c/\nu$ and $|\Delta\lambda/\lambda| = |\Delta\nu/\nu|$ and $\Delta s_m = c \cdot \Delta T_m$ the relation (11.25a, 11.25b) can be written as

$$\frac{\nu}{\Delta\nu} \leq \frac{\nu \cdot \Delta s_m}{c} = \nu \cdot \Delta T_m \tag{11.25b}$$

this gives the important relation

$$\Delta\nu \cdot \Delta T_m \geq 1. \tag{11.26}$$

For every spectrometer and interferometer the product of the minimum resolvable frequency interval $\Delta\nu$ and the maximum transit time difference ΔT_m between the interfering partial waves is equal to 1.

In order to increase the spectral resolution one has to increase the maximum path difference between the interfering partial waves. However, this is only possible up to an upper limit determined by the coherence length of the incident radiation (see Sect. 10.1).

Example

1. A general example of a common spectrometer or interferometer:
   $\Delta s_m = 1\text{m}$, $c = 3 \cdot 18^8$ m/s, => $\Delta T_m = 3.3$ ns, $\Rightarrow$ $\Delta\nu = 3\ 10^8\ \text{s}^{-1}$. For visible light ($\nu = 5\ 10^{14}\ \text{s}^{-1}$). This corresponds to a spectral resolution
   $$\frac{\nu}{\Delta\nu} = 1.7 \cdot 10^6.$$
2. For a grating spectrograph is $\Delta s_m = N \cdot d \cdot (\sin\alpha + \sin\beta) = N \cdot m \cdot \lambda \Rightarrow \lambda/\Delta\lambda = m \cdot N$. For $m = 1$ and $N = 10^5 \Rightarrow \Delta\lambda = 10^{-5}\lambda = 5 \cdot 10^{-3}$nm for $\lambda = 500$ nm.
3. For the Fabry-Perot interferometer is $\Delta s_m = 2F^* d \cdot \cos\alpha$ (see Eq. 10.32) where F^* is the effective number of interfering partial waves. With $2d\cos\alpha = m \cdot \lambda \Rightarrow \Delta s_m = F^* \cdot m \cdot \lambda$. With $F^* = 100, m = 2 \cdot 10^5 \Rightarrow \lambda/\Delta\lambda = 2 \cdot 10^7$.

Summary

- The angular resolution of the human eye is limited by diffraction and by the distance between the photoreceptors. The minimum still resolvable angle is

$$\varepsilon_0 = 3 \cdot 10^{-4}\text{rad} \approx 1'.$$

- The magnification V of an optical instrument is defined as V = (visual angle ε with instrument/visual angle ε_0 without instrument) where $\varepsilon_0 = G/s_0$ is the visual angle under which the object G at the visual range $s_0 = 25$ cm appears.
- A lens generates for a given image distance b a distinct image only for a finite interval Δa of the object distance a. The blurring of the image is within this object distance interval still smaller than the spatial resolution of the eye. This range Δa is called focus depth. It increases with decreasing diameter of the entrance pupil.
- The minimum possible angular resolution $\delta_{\min}$ of an optical instrument is limited by diffraction. With a diameter D of the imaging lens is $\delta_{\min} \geq 1.22\,\lambda/D$. The angular resolving power is defined as the inverse $R = 1/\delta_{\min} = D/(1.22\ \lambda)$.
- A classical microscope can only resolve structures of an object that are larger than $\lambda/2$.
- The imaging of an object through an optical system is only possible, if at least the zeroth and the first diffraction orders are transmitted by the system. (Abbe's theory of imaging). The zeroth diffraction order alone cannot produce a recognizable image.
- The transmission of optical systems is limited by the minimum cross section common to all incident light beams (entrance pupil) and that of the light beams on the image side (exit pupil). The measure for the luminosity of a lens with diameter D is the acceptable solid angle $\Omega = (\pi/4)(D/f)^2$.
- Spectrometers are optical instruments that are based on the refraction or diffraction of light by prisms or gratings, which result in a spatial separation of the different wavelengths.
- Interferometers are based on the separation of the incident wave into two or many partial waves which traverse different path lengths and are then superimposed. The interference of these partial waves depends on the wavelength λ and is used for the precise determination of λ.
- The spectral resolving power of any spectral apparatus

$$\frac{\lambda}{\Delta\lambda} = \frac{\Delta s_\text{m}}{\lambda}$$

is equal to the maximum path difference Δs_m between the interfering partial waves measured in units of the wavelength λ.

- For the prism spectrograph is

$$\frac{\lambda}{\Delta\lambda} = \frac{a}{2}\frac{\text{d}\theta}{\text{d}\lambda},$$

where a is the diameter of the entrance pupil and $\text{d}\theta/\text{d}\lambda \propto \text{d}n/\text{d}\lambda$ the angular dispersion which depends on the spectral dispersion of the prism material with refractive index $n(\lambda)$.

- For the grating spectrograph is

$$\frac{\lambda}{\Delta\lambda} \leq m \cdot N$$

dependent on the interference order m and the number N of illuminated grating grooves.

Problems

11.1 A lens generates a sharp image of the sun on a screen 2 m away from the lens. How large are the focal length f, the diameter d of the sun image and the lateral magnification? Which angular magnification is achieved with this lens, if the sum image is observed at the visual range s_0 = 25 cm?

11.2 A magnification glass with f = 2 cm is placed at a distance a = 1.5 cm above a book page in order to view the magnified letters. The eye of the observer is accommodated on the distance to the virtual image of the letters. What is the angular magnification? How large appears a letter with 0.5 mm size to the observer?

11.3 Derive, analogue to the derivation of Eq. (9.26) the more general Eq. (11.2).

11.4 The two components of a double star system have the angular distance $\varepsilon = 1.5''$. What is the minimum diameter of a telescope outside of our atmosphere which resolves both components? What is the minimum angular distance between two stars which can be still resolved by the naked eye?

11.5 What is the visual angle ε_0 under which the diameter of Jupiter appears to the naked eye? Why do planets not twinkle contrary to stars?

11.6 Sometimes one can read in newspapers that a telescope on board of a satellite in the altitude $h = 400$ km above ground can recognize a tennis ball ($d = 10$ cm) on earth. Is this possible? How large should be the diameter of the telescope, if air turbulence is neglected? Which is the minimum resolvable size of objects on earth, if the air turbulence is taken into account?

11.7 A radar system operating at λ = 1 cm should resolve structures of 1 m on an object 10 km away. Which angular resolution is necessary and what is the minimum size of the parabolic antenna?

11.8 A fine wire grating with a wire distance $d = 20\,\mu$m is viewed with relaxed eye (i.e. accommodated to $a = \infty$) through a microscope. The objective lens of the microscope has the angular magnification V = 10. Which focal length f_2 of the ocular lens has to be chosen that the grating wires appear to the eye like a mm-scale?

11.9 An optical diffraction grating ($d = 1\,\mu$m, size 10 × 10 cm) is illuminated by light with wavelength $\lambda = 500$ nm under the angle $\alpha = 60°$. What is the distance of two slit images $S_1(\lambda_1)$ and $S_2(\lambda_2)$ in the observation plane of a grating spectrograph with $f_1 = f_2 = 3$ m for $\lambda_1 = 500$ nm and $\lambda_2 = 501$ nm? How large is the bottom width of the zeroth diffraction maximum for an entrance slit width b → 0? What is the maximum width b of the entrance slit for the resolution of the two wavelengths?

11.10 (a) What are the spectral resolving power and the free spectral range of a Fabry-Perot interferometer with a plate separation $d = 1$ cm and a reflectivity of the mirrors R = 0.98

(b) In order to achieve an unambiguous assignment of a wavelength λ with this interferometer a prism spectrograph is placed before the FPI. What should be its focal length f to achieve the total separation of two wavelengths with a distance $\Delta\lambda$ equal to the free spectral range of the FPI when the slit width is b = 10 μm and the dispersion $dn/d\lambda = 5000/\text{cm}$?

References

1. D. Atchison, G. Smith: Optics of the Human Eye. (Butterworth-Heinemann, 2000)
2. C.W. Oyster; The Human Eye: Its Structure and Function (Sinauer, 2006)
3. Cario Cavalotti Luciano Certulli ed. :Age Related Changes of the Human Eye. (Humana Press 2008)
4. J. Mertz: Introduction to Optical Microscopy (Roberts and Company Publisher 2009)
5. H. Charles King: The history of the telescope (Dover Publications 2003)
6. https://en.wikipedia.org/wiki/Very_Large_Telescope
7. P. Jacquot: Speckle Interferometry, A review of the principal methods. (Strain 44, 57, (2008, Blackwell Ltd.)
   R. Dändliker, and P. Jacquot, (1992) Holographic Interferometry and Speckle Techniques. Optical Sensors. VCH Verlagsgesellschaft, Weinheim, 589–628
8. W. Demtröder: Laser Spectroscopy Vol. 1 5th edit. (Springer Heidelberg 2014)
9. S.W. Hell et. al.: Nanoscale Resolution with Focused Light. In: Handbook on Confocal Microscopy ed. by J. Pawley (Springer Heidelberg 2006)
10. Hell, S. W. and Wichmann, J. Optics Letters **19**, 780–782 (1992) Osseforth, C., Moffitt, J. R., Schermelleh, L. and Michaelis, J. Optics Express (2014), **22**, 7028–7039
11. H. Erfle: Super Resolution Microscopy: Methods and Protocols (Human Press 2017)
12. John James: Spectrograph Design Fundamentals Cambridge Univ. Press (2012)

New Techniques in Optics

12

For the last years several new optical techniques have been developed and applied. Some of them are based on old ideas but could not be realized because the technical requirements were missing, while also some completely new concepts have led to astonishing results. They have meanwhile often captured many fields in daily life and have enlarged considerably the possibilities of optics and its applications.

In this chapter we will present some of these techniques, which will illustrate that we are experience the beginning of an "optical revolution".

The references at the end of the chapter give the possibility for a more detailed information about the different techniques.

12.1 Confocal Microscopy

The confocal microscopy combines the high radial spatial resolution (in the x, y-plane perpendicular to the light propagation) with a comparable high resolution in the z-direction (in the direction of the optical axis). A confocal microscope can be regarded as an instrument with extremely short focus depth. It has an essentially better suppression of stray light than the classical microscope. Its principal design is illustrated in Fig. 12.1. The light from a light source (here generally lasers are used because of the necessary higher intensity) is focused onto a small circular aperture B_1, is reflected by a beam splitter BS and focused by the lens L_2 onto the plane $z = z_0$ of the object under investigation. The backscattered light is collected by L_2, focused after transmission through BS onto a pinhole aperture B_2 and reaches the detector. Light scattered from other planes $z = z_0 \pm \Delta z$ of the object is not focused onto the pinhole and therefore only a small fraction reaches the detector. The pinhole B_2 therefore suppresses light from other planes in the object. The maximum signal is obtained for light that is exactly focused onto the pinhole. Because a focal spot of the object is imaged onto the focus in the pinhole B_2 the technique is called "**confocal microscopy**" [1]. When shifting the object into the z-direction other z-planes of the object are selectively viewed. Shifting the sample in the x, y-plane yields z-dependent images for any point in the x, y-plane which gives three-dimensional images of the sample. This is used in cell biology to win three-dimensional structures of cells. They are generated by a computer program that calculates the structure from the z-dependent transmitted intensity of the backscattered light.

The spatial resolution in the x, y-planed is given by the size of the pinhole, the distance of the plane $z = z_0$ from L_2 and the focal length f_2 of L_2. Instead of shifting the sample one may also use a turnable mirror as shown in Fig. 12.2. Turning the mirror about the y-axis gives a scan of the sample along the x-axis. The backscattered light is always focused onto the pinhole B_2 [2, 3], The measurement can be also performed at daylight, if a spectral filter is placed before

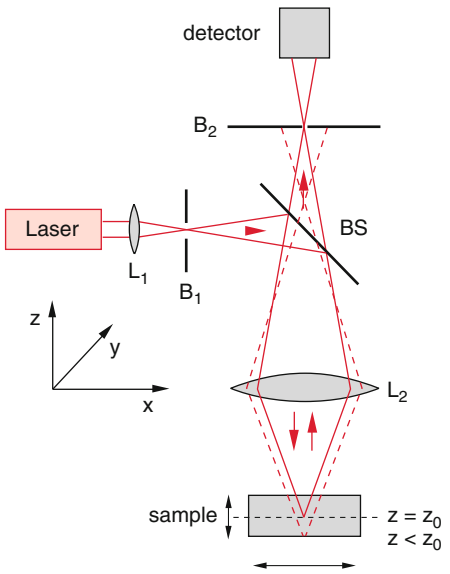


Fig. 12.1 Confocal microscopy

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4_12

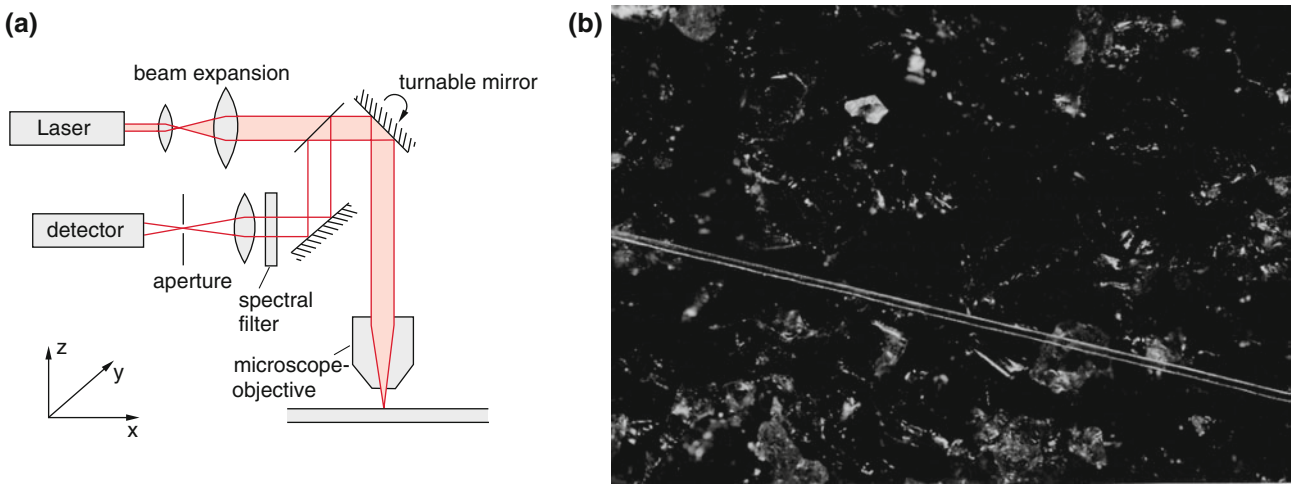


Fig. 12.2 Confocal microscopy with raster principles using laser beams **a**) experimental arrangement **b**) measurement of the rough surface of a grinding disc with a human hair for size comparison, (H. Jochen Foth, TU Kaiserslautern)

the detector to restrict the detected light to the narrow spectral interval of the illuminating laser.

The main application fields are

- Fluorescence microscopy of biological cells where selected parts of the cell are illuminated by the focused laser beam and the resulting fluorescence is detected with high spatial resolution.
- Investigation of surface structures and molecules adsorbed on the surface. As example the confocal microscopy of a grinding disc is shown in Fig. 12.2b, where a human hair is included for size comparison.

An interesting version of confocal microscopy is shown in Fig. 12.3, where larger areas of the sample can be viewed simultaneously. The widened beam of a laser is sent through a perforated mask (this is an opaque screen with many small holes) and reflected by the beam splitter onto the sample. In the sample plane a pattern of focal points is generated which are the source of scattered light or fluorescence. They are imaged by the lens L onto a CCD-camera. The perforation pattern is designed in such a way that it corresponds to the pixels of the CCD-camera. The output of the CCD camera gives a picture of all illuminated points of the sample and measures therefore a larger area of the sample simultaneously with high spatial resolution.

In Fig. 12.4 chromosomes of a human cell are shown, which are made visible by transmitted light using differential interference contrast techniques. The bright chromosomes were selectively excited by the illuminating light and their fluorescence was spatially resolved by confocal microscopy [4].

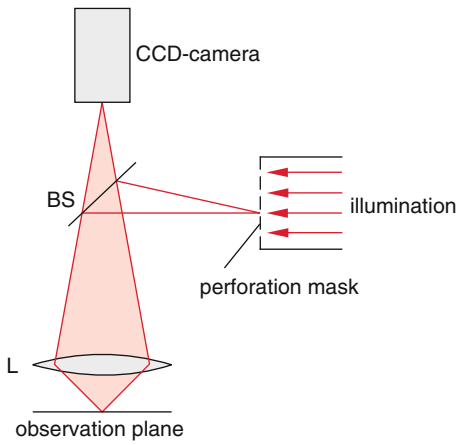


Fig. 12.3 Confocal microscopy with shadow masc and CCD camera

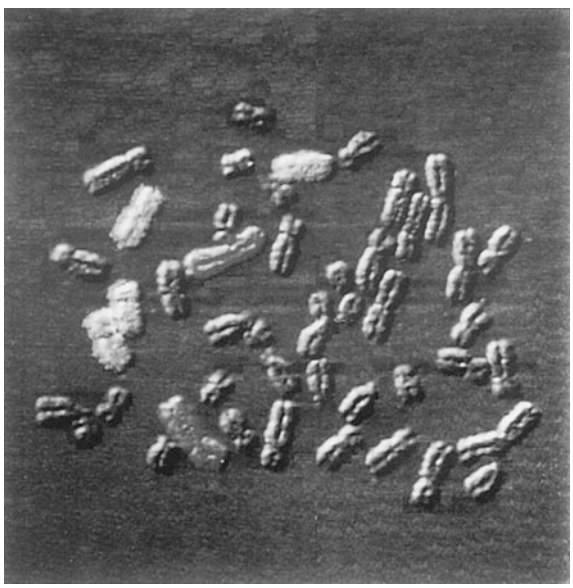


Fig. 12.4 Confocal microscopy of chromosomes in the human cell [4]

12.2 Optical Near Field Microscopy

In Sect. 113.3 we have learned that with a classical microscope structures below $\lambda/2$ cannot be resolved when illuminated by light with the wavelength λ. This resolution limit due to diffraction can be surpassed by optical near field microscopy [5] which is especially used for the investigation of tiny structures on surfaces. Its principle is illustrated schematically in Fig. 12.5.

The surface is illuminated by an intense laser. The light scattered by structures on the surface is detected behind a very small aperture (about 100 nm diameter) which is brought close to the surface (Fig. 12.5a). Shifting the aperture at a constant distance z from the surface across the surface the scattered light intensity $I(x, y)$ is measured as a function of the location (x, y) on the surface. This gives information about the structure of the surface which influences the scattered light intensity.

This method is a raster scanning technique, where the information about the different points (x, y) of the surface is not obtained simultaneously but sequentially in time. A computer converts the measured signals into a three-dimensional picture of the surface, if a model for the relation between scattered light intensity $I(x, y)$ and surface structure has been fed into the computer [5].

Often the light is transmitted through an optical fiber with the fiber-end placed closely above the surface (within a few nm). In order to increase the spatial resolution the fiber end is sharpened to a cone and the sides of this cone are covered by a thin metal layer which prevents that light leaves through these sides. (Fig. 12.5b). The light can then only escape through the narrow peak (diameter 50 nm) and illuminates a small spot on the surface. The light scattered from this tiny spot is collected by a lens and focused onto the detector. With this technique a resolution of 20–30 nm can be reached which surpasses that of conventional microscopy by one order of magnitude.

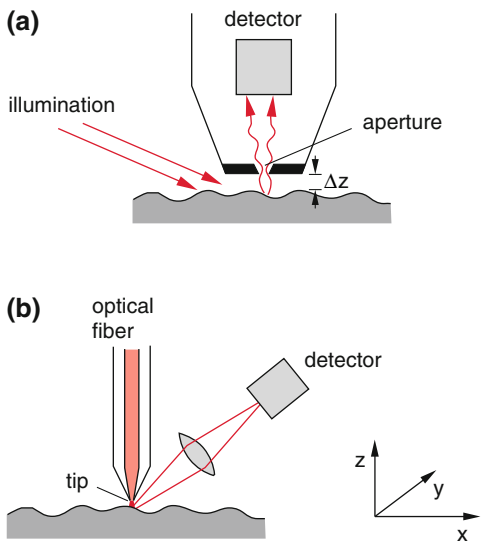


Fig. 12.5 Near field microscopy: spatial resolution of structures $\Delta x < \lambda/2$ on surfaces **a**) by measurements of laser light scattered from the illuminated surface and collected through a very small aperture ($d \ll \lambda$) close to the surface onto the detector **b**) illumination of a small spot on the surface through the sharpened peak of an optical fiber closely above the surface and detection of light scattered from the surface

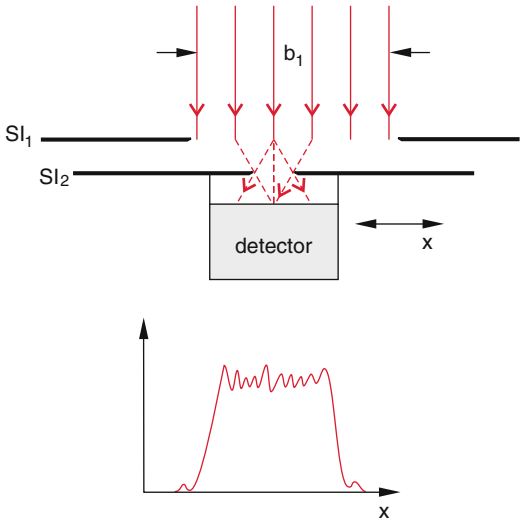


Fig. 12.6 Optical near field microscopy of the Fresnel diffraction closely behind the slit

With increasing spatial resolution the light power, received by the detector drastically decreases. One has therefore to use high light powers of the laser and sensitive detectors with a small background noise.

Since with this near field method the scattered light in the near field of the scattering centers is detected where Fresnel diffraction governs the total field amplitude (see Sect. 10.6) the interpretation of the results is not straight forward and demands sophisticated computer programs. This can be seen for the simple example of an illuminated slit with width b_1 (Fig. 12.6). When a second slit with width $b_2 < b_1$ is shifted across the transmitted light at a distance $d < b_1$ the near field of the Fresnel diffraction is detected and the measured intensity $I(x)$ shows a complicated structure which sensitively depends on the distance d from the first slit because the phase differences between the different diffracted partial waves strongly change with the distance d.

More information about the near field microscopy can be found in [6–8] and in the Journal *Scanning Microscopy*.

12.3 Active and Adaptive Optics

For astronomical telescopes (mirror telescopes see Sect. 11.2.3) the light power received from objects at far distances is proportional to the area of the primary mirror.

Therefore the mirror should be as large as possible. For very large parabolic mirrors (diameter 6–10 m) the mirror cannot be made thick enough (because of weight problems) to completely avoid varying bending when the telescope is oriented towards different celestial objects. This bending changes the parabolic surface of the mirror and deteriorates the imaging quality. Thicker mirrors would extremely increase the production and transportation costs and would also demand much stronger mirror mounts.

The solution of this problem is the active optics.

12.3.1 Active Optics

The mirror is kept as thin as its stability allows and can therefore bend under the influence of gravitational forces. This bending is, however, avoided by many adjustable control elements in form of piezo-rods which are mounted on the backside of the mirror (Fig. 12.7). These rods change their length depending on the voltage applied to the piezo-element. They are controlled by a computer in such a way, that the mirror surface always forms the wanted rotational paraboloid, independent of the mirror position. For illustration Fig. 12.7b shows the backside of the 8 m mirror of the very large telescope VLT on the mountain *Paranal* in Chile with the many control elements.

This technique of active optics is nowadays used for all modern large telescopes. For very large mirrors (for instance the 10 m mirror of the Keck telescope on the Mouna Kea on Hawai completely new techniques have been invented. Many small mirrors are united to a large mirror (Fig. 12.8) in such a way that their surfaces are perfectly aligned to form the parabolic surface of a large mirror which images the star light into the common focus This implies that the relative position of these many small mirrors has to be aligned within $\lambda/10$ and has to be stable when the telescope is turned to the wanted direction towards a star under investigation. For such a honey comb telescope the production and transportation costs are much lower, but the assembling and alignment of up to 100 single mirrors to a large mirror demands high technical efforts and knowledge [9, 10].

Several large mirrors can be optically combined by optical beamlines and act as huge interferometers [11] with a spatial resolution that exceeds by far that of single mirrors. In such arrangements the light collected by the different mirrors is transported by optical beam lines to a common location where the different contributions superimpose and interference structures can be observed which strongly depend on the location of the light source (i.e. the star). The optical path lengths Δs_i between the light from the different

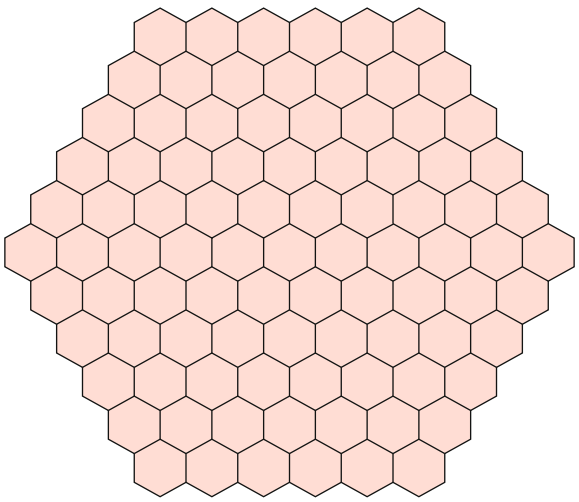


Fig. 12.8 Primary mirror of an astronomical telescope which is composed of many hexagonal segments. The whole surface forms a paraboloid with maximum deviations $<\lambda/10$ from the ideal surface

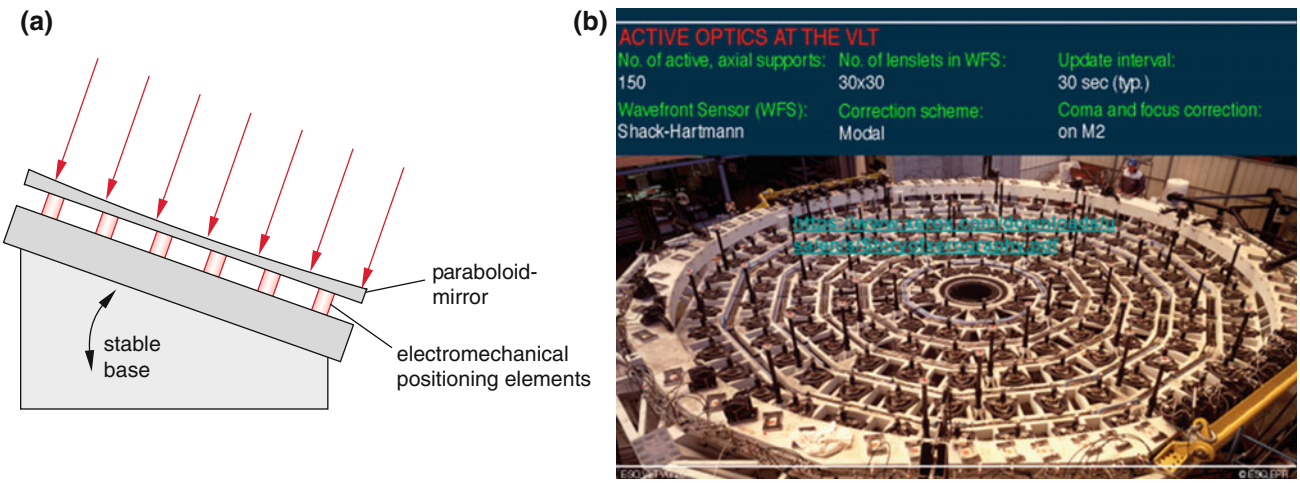


Fig. 12.7 Paraboloid mirror with active optics **a**) principle design **b**) backside of the large primary mirror with the many piezo-electric positioning elements (Very large telescope of the ESO at the Paranal in Chile)

telescopes should not fluctuate by more than $\lambda/10$ for path length of many meters. For a diameter D of the total telescope array the diffraction limited angular resolution can be decreased by this technique to $\Delta\varepsilon \approx \lambda/D$. With a realistic number of $D = 200\,\text{m}$ this gives for $\lambda = 1\,\mu\text{m}$ the angular resolution $5 \times 10^{-9}\,\text{rad} = 0.001''$.

This is quite similar to the diffraction by an optical grating with N grooves, a groove distance d and a total width $D = Nd$. Here the diffraction limited width of the interference maximum is the same as that caused by a slit with width D.

12.3.2 Adaptive Optics

The angular resolution of large ground based telescopes does not reach by far the diffraction limit due to turbulences in the atmosphere and uprising air caused by local thermal heating. This leads to time-dependent changes of the refractive index resulting in a fluctuation of the deflection of star light.

The image of a star in the observation plane of a telescope moves randomly around a center and yields for longer illumination times an average intensity distribution which has a much larger diameter than the diffraction limited image [12].

Example

For a telescope with mirrors diameter $D = 1$ m the diffraction limited angular resolution is for $\lambda = 500\,\text{nm}$ $\Delta\varepsilon_{\text{diffr}} \approx \lambda/D \approx 5 \times 10^{-7}\,\text{rad} = 0.1''$. However, the air turbulence limits the angular resolution to about $\Delta\varepsilon \approx 1''$ i.e. to 10 times the diffraction limit. This resolution limited by air turbulence is called by astronomers "*seeing*". For larger telescopes the difference between diffraction limited resolution and air turbulence limited resolution becomes accordingly larger.

On high mountains this effect becomes smaller because the path length through the atmosphere is shorter and water droplets or dust particles are much less abundant. Nevertheless also for selected locations of telescopes the air turbulence still limits the angular resolution. This is shown in Fig. 12.9. which compares the diffraction limited image in the observation (x, y)-plane with the broadened image caused by air turbulence. Instead of the intensity distribution $I(r) = \sin^2 r/r^2$ $(r^2 = x^2 + y^2)$ of the diffraction intensity maximum one obtains a more or less irregular intensity distribution over a much larger area.

The perturbations caused by air turbulence can be greatly reduced by the technique of **adaptive optics**. The technique works as follows (Fig. 12.10).

The star light reflected by the secondary mirror M_2 is formed by the lens L_1 into a parallel beam which falls onto a

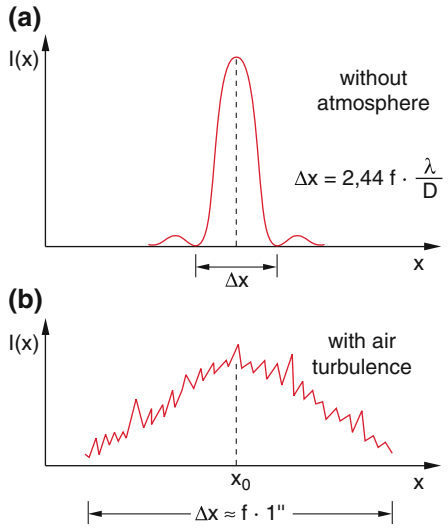


Fig. 12.9 Intensity distribution of the star image **a**) diffraction limited distribution without distortion by the atmosphere. **b**) Speckle picture, broadened by refractive index fluctuations in the atmosphere

flat thin mirror M_3 which has control elements on its backside which can deform the mirror surface (active optics). The light reflected by the deformable mirror is reflected by mirror M_4 passes through a beam splitter BS. The reflected light is detected by a wave front sensor which detects any deformation from a plane wave front, while the light transmitted by BS reaches the detector and forms the image of the observed star. The wave front sensor delivers a signal which is proportional to the deviation from a plane wave front. This signal is sent to the actuators on the backside of the deformable mirror M_3

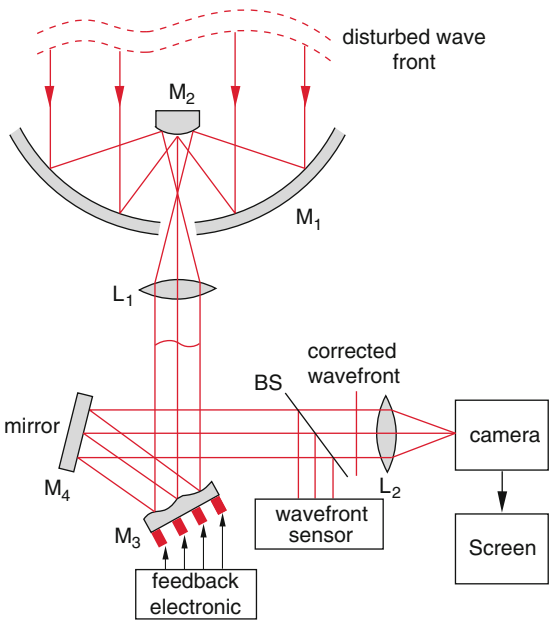


Fig. 12.10 Principle of adaptive optics

which deform M_3 in such a way that the wave fronts become planes and the image of the star becomes nearly diffraction limited. Since the air turbulence can change within a fraction of a second the feedback control must be fast enough. Therefore it needs a sufficiently large input signal, i.e. a bright radiation source. Since such bright stars are not available for all positions of the telescope an artificial star is created by sending a laser beam along the axis of the telescope with a laser wavelength tuned to the yellow absorption line of sodium atoms. In the altitude of about 50 km a layer of sodium atoms exists. The excited sodium atoms emit fluorescence which represents a very bright light source that is used for activating the feedback control [12]. In Fig. 12.11 the effect of the adaptive optics is illustrated by exposures of the star Cygnus α with and without adaptive optics.

The adaptive optics can be, of course, also used for observing objects on earth with a telescope.

Special techniques of nonlinear optics (four wave mixing) have been developed which correct distorted wave fronts of incident light after reflection by mirrors of selected materials (liquids or gases) (phase conjugated mirrors) [13].

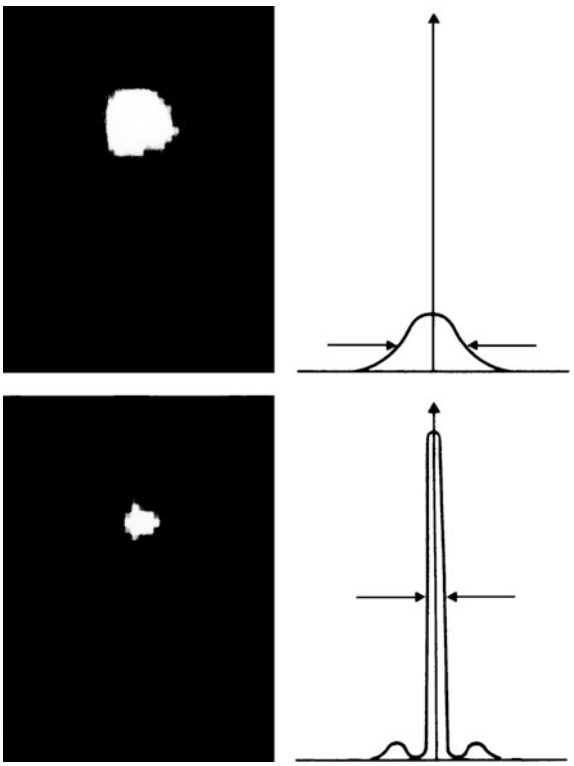


Fig. 12.11 Image of the star Cygnus α **a**) without and **b**) with adaptive optics

12.3.3 Interferometry in Astronomy

The fundamental limit for the angular resolution $\Delta\varepsilon > 1.22\lambda/D$ of telescopes with diameter D is given by diffraction. For a diameter $D = 5$ m of the primary mirror this implies for $\lambda = 1$ μm a limit of $\Delta\varepsilon > 2.4 \times 10^{-7}$ rad $= 0.05''$ if the effect of air turbulence has been eliminated by adaptive optics. Since 2005 interferometric methods can be applied where two or more telescopes with a distance $\Delta x \gg D$ can be connected by optical beam lines in such a way that a coherent superposition of the signals from the two telescopes is achieved. This results in an interference structure which depends on the phase difference between the signals from the two telescopes (Fig. 12.12).

The angular resolution of this coupled system corresponds to that of a telescope with a mirror diameter Δx and is therefore with $\Delta x \gg D$ much higher than that of a single telescope although the received light power is only twice as large. The technical challenge is tremendous, because the light path from the telescopes to the detector, which can be longer than 100 m, has to be kept constant within $\lambda/10$. This can be only achieved with an electronic feedback for the control of the path length. In order to change continuously the path difference for the two signal one of the two signals is sent to a retro-prism on a wagon (Fig. 12.13) which can move on precisely aligned smooth rails [11].

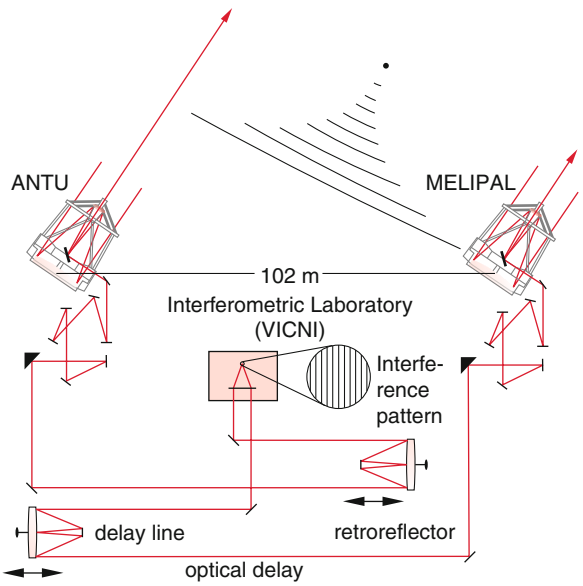


Fig. 12.12 Principal design of the interferometric detection with two large telescopes and adaptive optics at the Paranal in Chile (with kind permission of Dr. Glindemann [18])

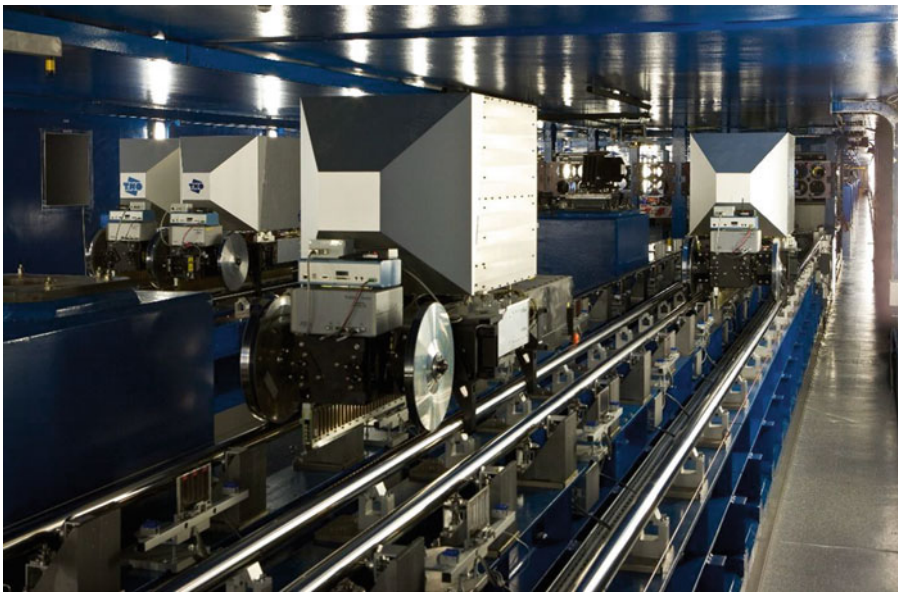


Fig. 12.13 Part of the interferometric path with a retro-reflecting mirror mounted on a waggon which moves smoothly on a track (ESO observatory on the Paranal in Chile)

12.4 Holography

Classical photography images an illuminated object through a lens system into the detector plane (Fig. 12.14a). Each point of the object is mapped into the corresponding point of the image. The detector signal is proportional to the incident intensity. Any information about the phase of the incident light is lost. This implies that no direct information about the three-dimensional structure of the object is obtained. The three-dimensional object is reduced to a two-dimensional image. The fact that we can recognize the three-dimensional object by looking at the two-dimensional image is due to our brain which can reconstruct the real object by comparison with earlier stored information.

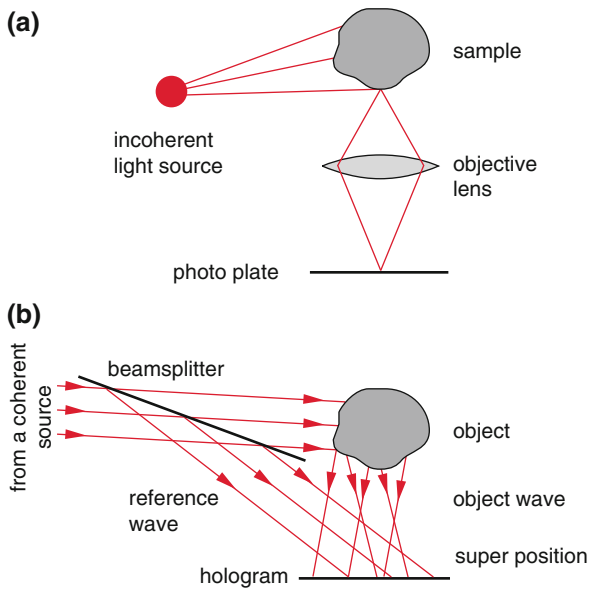


Fig. 12.14 Comparison between **a**) classical photography and **b**) holography

Denis Gabor (1900–1979) published 1948 the idea, to superimpose in the detection plane two coherent waves, namely the illumination wave scattered by the object, and a reference wave from the same source. This superposition generates an interference pattern on the photo plate in the detector plane, that contains information about the amplitude and the phase of the wave scattered by the object. This allows the determination of the distance between the different object points and the photo plate (Fig. 12.14b). The blackness pattern on the photo plate is called a *hologram*. After developing the photo plate and illuminating it with light of the same wavelength a three-dimensional picture of the object is "reconstructed". This idea laid the foundation of a completely new optical technique, called **holography**. Gabor received for his invention the Nobel Prize 1971.

Since this method demanded coherent light sources with high intensity, Gabor could realize his idea merely imperfectly. Only after the development of lasers (see Vol. 3, Chap. 7) holography started its triumph [14–16].

12.4.1 Recording of a Hologram

In Fig. 12.15 the principle of the recording of a hologram is illustrated schematically. The output beam of a laser is widened by a lens system and split into two beams by a beam splitter. The reflected wave is the reference wave

$$\boldsymbol{E}_0 = \boldsymbol{A}_0 \mathrm{e}^{\mathrm{i}(\omega t - \boldsymbol{k}_0 \cdot \boldsymbol{r})} \tag{12.1}$$

which is sent to the photo plate in the x-y-plane. The transmitted partial beam illuminates the object. The wave scattered by the object has the amplitude in the plane of the photo plate

$$\boldsymbol{E}_\mathrm{s}(x,y) = \boldsymbol{A}_\mathrm{s} \mathrm{e}^{\mathrm{i}[\omega t + \varphi_\mathrm{s}(x,y)]}, \tag{12.2}$$

where the phase $\varphi(x,y)$ depends on the distance of the object points that scatter the wave. The amplitude $E_\mathrm{s}(x,y)$ is the sum of all amplitudes which are scattered by the different object points into the point (x,y) on the photo plate. Also the phase $\varphi(x,y)$ in the point (x,y) is determined by the superposition of the phases of all partial waves scattered from all illuminated points of the object. The total intensity at the point $\boldsymbol{r}_0 = \{x,y,0\}$ is then

$$\begin{aligned} I(x,y) &= c\varepsilon_0 |E_\mathrm{s}(x,y) + E_0(x,y)|^2 \\ &= c\varepsilon_0 \big| A_0^2 + A_\mathrm{s}^2 + A_0^* A_\mathrm{s} \mathrm{e}^{\mathrm{i}[\boldsymbol{k}_0 \cdot \boldsymbol{r}_0 - \varphi_\mathrm{s}(\boldsymbol{r}_0)]} \\ &\quad + A_0 A_\mathrm{s}^* \mathrm{e}^{-\mathrm{i}[\boldsymbol{k}_0 \cdot \boldsymbol{r}_0 - \varphi_\mathrm{s}(\mathbf{r}_0)]} \big| \\ &= c\varepsilon_0 |A_0^2 + A_\mathrm{s}^2 + 2A_0 A_\mathrm{s} \cos(\varphi_0 - \varphi_\mathrm{s})|, \end{aligned} \tag{12.3}$$

where $\varphi_0 = \boldsymbol{k}_0 \cdot \boldsymbol{r}_0$. The phase difference $(\varphi_0 - \varphi_\mathrm{s})$ depends on the optical path difference between reference- and scattered wave. The phase dependent interference term in (12.3) contains the desired information about the distance of the different object points from the points (x,y) on the photo-plate i.e. they give the three-dimensional structure of the object.

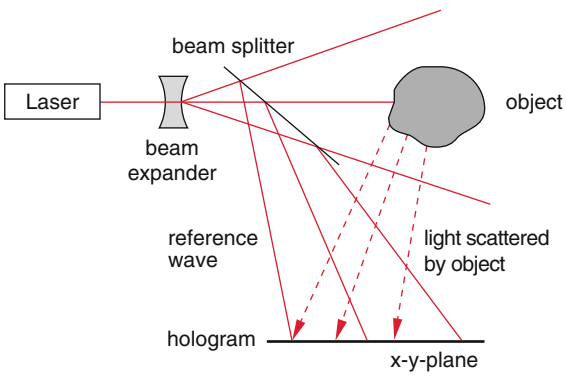


Fig. 12.15 Optical setup for taking a hologram

Example

1. We assume that the object is a plane illuminated by a plane wave and reflects this wave (Fig. 12.16). In the plane of the photo plate is $\boldsymbol{k}_0 \cdot \boldsymbol{r}_0 = k_0 x \cdot \sin\alpha_1$; $\varphi_\mathrm{s} = -k_0 \cdot x \sin\alpha_2 \Rightarrow (\varphi_0 - \varphi_\mathrm{s}) = k_0 x(\sin\alpha_1 + \sin\alpha_2)$. The cosine-function in (12.3) has the period $\Delta x = \lambda/(\sin\alpha_1 + \sin\alpha_2)$. The superposition of reference and object wave results in a periodic intensity modulation in x-direction on the photo plate with a distance between the intensity maxima

$$d = \frac{\lambda}{\sin\alpha_1 + \sin\alpha_2},$$

which depends on the angles α_1 and α_2 between the normal of the two waves and the surface normal of the photo plate. On the developed photo plate occurs a periodic pattern of parallel stripes with a sinewave distribution of the blackening.

Such a periodic blackening pattern can be used as holographic transmission grating with the grating constant d. When the grating is illuminated by a plane wave from the same laser, one obtains for the correct choice of the angle α a plane wave with the phase plane coincident with the object plane.

If the illuminated positions on the photographic layer are removed by chemical techniques with subsequent etching and coating with a reflecting layer a holographic reflection grating can be produced. Such gratings have a perfect grating constant d without any errors because $d = \Delta x$ is determined by the optical

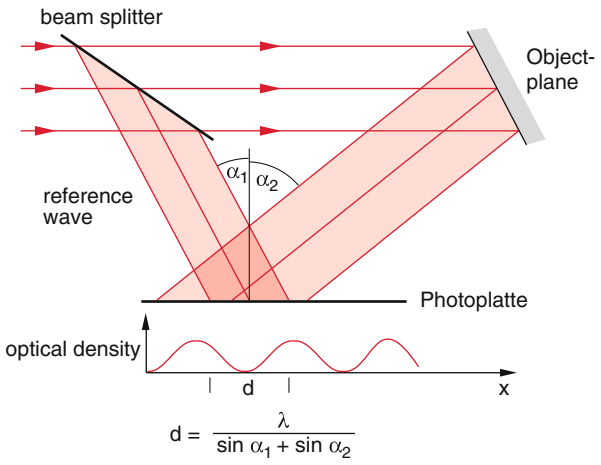


Fig. 12.16 Production of a holographic grating by superposition of two plane waves where the wave vectors form the angles α_1 and α_2 against the normal of the grating plane

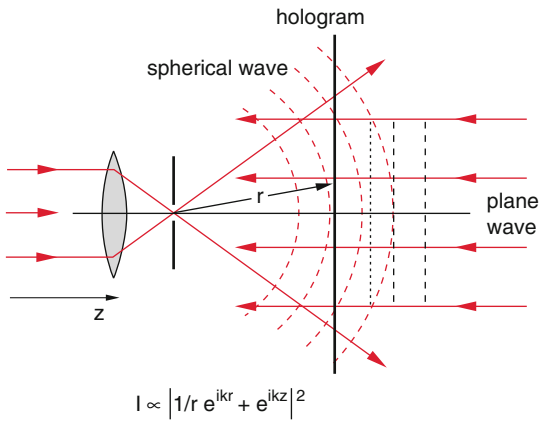


Fig. 12.17 The superposition of a plane wave with a spherical wave results in a circular interference ring system. The corresponding hologram represents a Fresnel zone plate

wavelength λ of the illuminating laser light used for the production of the grating. Their drawback is, however, the sinewave structure of the grooves in contrast to the ruled gratings which have a staircase shaped form. The reflectivity of holographic gratings is therefore smaller than that of ruled gratings and they have no blaze angle (see Sect. 10.5.2).

2. A plane wave superimposes a spherical wave (Fig. 12.17). The resulting hologram shows a structure of black circles, which corresponds to a Fresnel's zone plate. When the developed hologram is illuminated by a plane wave a spherical wave originates which is focused into a point that corresponds to the center of the spherical wave used for producing the hologram. The blackness of the photo plate is proportional to the incident intensity and the contrast between maximum and minimum blackening depends on the difference between the amplitudes of the two waves. If the amplitudes of the two waves have a ratio of 1:10 the contrast is

$$K = (I_{max} + I_{min})/(I_{max} - I_{min}) = (1.1/0.9)^2 = 1.5$$

Note: In classical photography each point of the object is imaged into a well-defined point of the image, while in holographic imaging the wave scattered by a point of the object is distributed over the whole hologram. This implies that each part of the hologram contains already information about the whole object. One can, for instance, cut the hologram into two pieces. Each piece can produce the whole three-dimensional object, although with lower quality.

12.4.2 The Reconstruction of the Wave Field

In order to obtain a three-dimensional image of the object from the hologram which contains the information about the object in encoded form (Fig. 12.18) the developed photo plate has to be illuminated by a coherent plane wave, the reconstruction wave

$$\boldsymbol{E}_r = \boldsymbol{A}_r \cdot e^{i(\omega t - \boldsymbol{k}_r \cdot \boldsymbol{r})} \tag{12.4}$$

with the same light frequency ω as the illuminating wave, used for the exposure of the hologram (Fig. 12.19). The amplitude of the reconstruction wave transmitted by the hologram

$$A_T = T(x, y) \cdot A_r \tag{12.5}$$

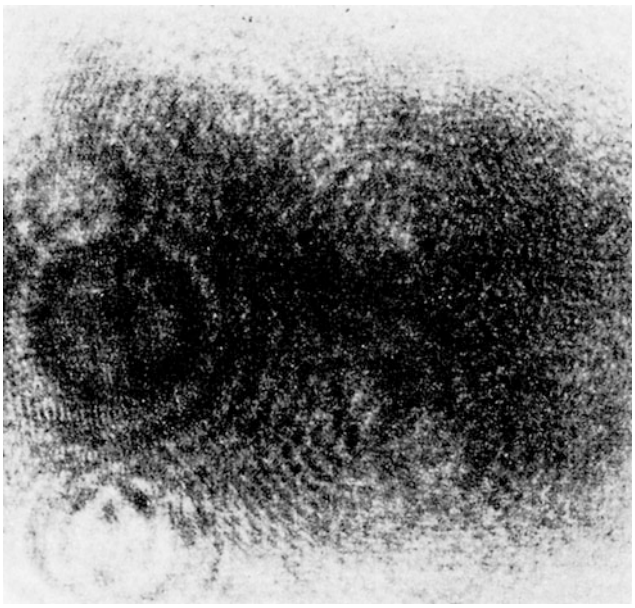


Fig. 12.18 Hologram of a chess-board (from: H. Nassenstein, Z. Angew. Physik 22, 37–50 (1966))

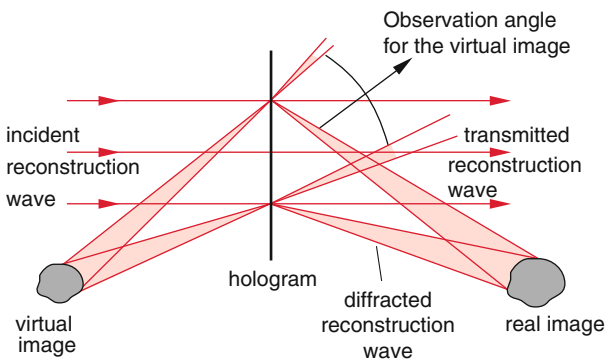


Fig. 12.19 Reconstruction of a hologram

depends on the blackening of the hologram which in turn is proportional to the intensity (12.3) used for the exposure. The transmission of the developed photo-plate is

$$T(x,y) = T_0 - \gamma I(x,y) \tag{12.6}$$

where γ is the blackening coefficient of the photo plate and $I(x,y)$ the intensity (12.3). The transmitted reconstruction wave

$$E_T = T(x,y) \cdot A_r \cdot e^{i(\omega t - k_r r)} \tag{12.7a}$$

has the amplitude

$$\begin{aligned} A_T = {} & A_r T_0 - \gamma A_r (A_0^2 + A_s^2) \\ & - \gamma A_r A_0^* A_s e^{i(k_0 \cdot r_0 - \varphi_s)} \\ & - \gamma A_r A_0 A_s^* e^{-i(k_0 \cdot r_0 - \varphi_s)} \end{aligned} \tag{12.7b}$$

The first two terms describe an attenuation of the transmitted wave that is independent of the location (x,y). The last two terms represent two new waves

$$E_{T_1} = -\gamma A_0^* A_r A_s e^{i[\omega t - (k_r - k_0) \cdot r_0 - \varphi_s]} \tag{12.8a}$$

$$E_{T_2} = -\gamma A_0 A_r A_s^* e^{i[\omega t - (k_r + k_0) \cdot r_0 + \varphi_s]} \tag{12.8b}$$

which propagate into the directions $\boldsymbol{k}_1 = \boldsymbol{k}_r - \boldsymbol{k}_0$ and $\boldsymbol{k}_2 = \boldsymbol{k}_r + \boldsymbol{k}_0$.

Note: The directions of these waves are not identical with the direction of the reconstruction wave. The reconstruction wave transmitted through the hologram is diffracted by the blackening structures of the hologram, which act like an amplitude grating.

Both waves carry information about the amplitude A_s and the phase φ_s of the wave, scattered by the object and used for the exposure of the hologram, because they contain the amplitude

$$E_s = A_s \cdot e^{i(\omega t - \varphi_s)} \quad \text{resp.} \quad E_s^* = A_s^* \cdot e^{-i(\omega t - \varphi_s)} \tag{12.8c}$$

which have been also used for the exposure of the hologram.

As is shown in Fig. 12.19 two images of the same object appear: A virtual image produced by the wave E_{T_1} which appears behind the hologram and a real image due to the wave E_{T_2}. This real image can be made visible if a screen is placed at the location of the real image. However, this projection on the screen gives, of course, only a two-dimensional image.

Looking through the hologram in the direction towards one of the two waves a three-dimensional image of the object appears to the eye, which is equal to a view of the real object seen from the location of the photo plate [17, 18].

Remark When using another wavelength λ_r for the reconstruction wave as the wavelength λ_s of the original wave scattered by the object, the reconstructed image appears magnified or scaled down by the factor λ_r/λ_s.

12.4.3 White Light Holography

The wide distribution and acceptance of holography was fostered by the invention of white-light holography, because here no laser is needed for the reconstruction of the image, but only an incoherent light source (e.g. a light bulb or the sun).

How can we understand this?

For the construction of the hologram, which does need a laser, a special design has to be chosen (Fig. 12.20). A thin photographic layer on a glass plate is illuminated from above by the enlarged beam of a laser (reference wave) and below from the light scattered b the object (object wave).

In the photographic layer the intense reference wave and the weaker object wave superimpose and generate and interference stripes of maxima and minima, which are essentially parallel to the surface of the photo plate and generate a layer structure of the blackening (see for instance Fig. 12.16). For a wavelength $\lambda = 0.6\,\mu m$ and a layer thickness of $10\,\mu m$ about 20 parallel blackness layer are formed, which correspond to the interference layers of maximum intensity.

When the developed photo plate is illuminated by light with the wavelength λ the light is partially reflected and the partial waves reflected by the different layers interfere. Their path difference is for an incidence angle $\Delta s = 2d \cdot \sin\alpha$ (Fig. 12.20b). Only for those wavelength λ with a path difference

$$2d \cdot \sin\alpha = m \cdot \lambda \qquad (m = 1, 2, 3\ldots) \tag{12.9}$$

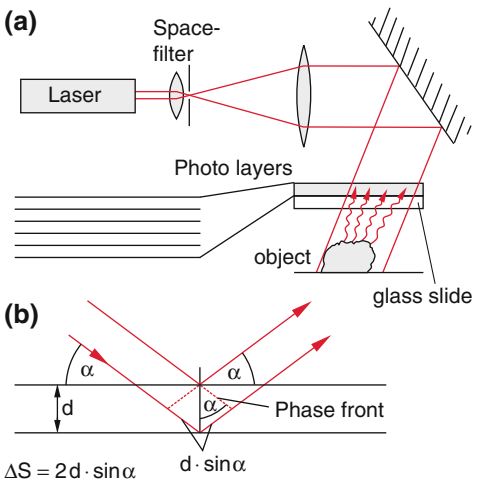


Fig. 12.20 White light holography **a**) taking the hologram **b**) selective reflection at the developed hologram due to interference between the partial waves reflected at the different layers

constructive interference occurs (Bragg condition). The layer structure generated at the construction of the hologram selects from the spectral continuum of the white light source only those wavelength that fulfill the condition (12.9). The reconstructed image of the object appears under white light illumination at a wavelength λ which depends on the incidence angle α. Changing the angle α also changes the color of the three-dimensional image.

12.4.4 Holographic Interferometry

In Sects. 10.3 and 10.4 some classical types of interferometers were introduced which are based on two-beam interference or multiple beam interference. They are used for the accurate measurement of small path differences Δs or of wavelengths λ_i of spectral lines.

Holographic interferometry broadens the possibilities of classical interferometer considerably and can be applied to many interesting areas of science and technology. There are in principle three procedures [19]:

(a) In the *real time technique* a hologram is recorded from an object at rest. The hologram-plate is now developed at the same place without moving it and is then illuminated with the reference wave, creating a hologram image in the same way as in Fig. 12.19. Now the object is exposed to external influences (such as pressure or temperature changes) and is illuminated by the same wave as before. The changes of the object appear as phase changes of the signal wave. The superposition of this altered signal wave with the reconstruction wave of the hologram before the object was modified, results in interference structures in the holographic image which appear only for such points of the object, which were modified. This technique allows the detection of object modifications that are much smaller than the wavelength λ. This is illustrated in Fig. 12.21a, which shows two exposures of a wine glass: At first a hologram of the wine glass is constructed and then the photo plate is developed and again a hologram is reconstructed. During the developing process the photo plate shrinks a little bit. Therefore the holographic image is a little bit larger. When this hologram is superimposed onto the first hologram, horizontal interference stripes are generated which give quantitative information about the magnitude of the shrinking.
If now the glass is filled with a hot lighter gas the rising gas changes the refractive index and causes deformed interference lines (Fig. 12.21b).

(b) In the double exposure method a hologram of the object is taken before the modification of the object takes place and then for a fixed position of the photo plate a second hologram on the same photo plate after the modification. For example, the deformation of a metal plate under the influence of external forces is measured by making a hologram before the deformation, then deforming the plate without moving it from its position and afterwards take a second hologram (double exposure of a fixed photo plate). In Fig. 12.22 a double exposed hologram of an aluminum disc is shown. The black lines give information about the magnitude of the deformation. The wave scattered by the object has a different phase before and after the deformation. For a

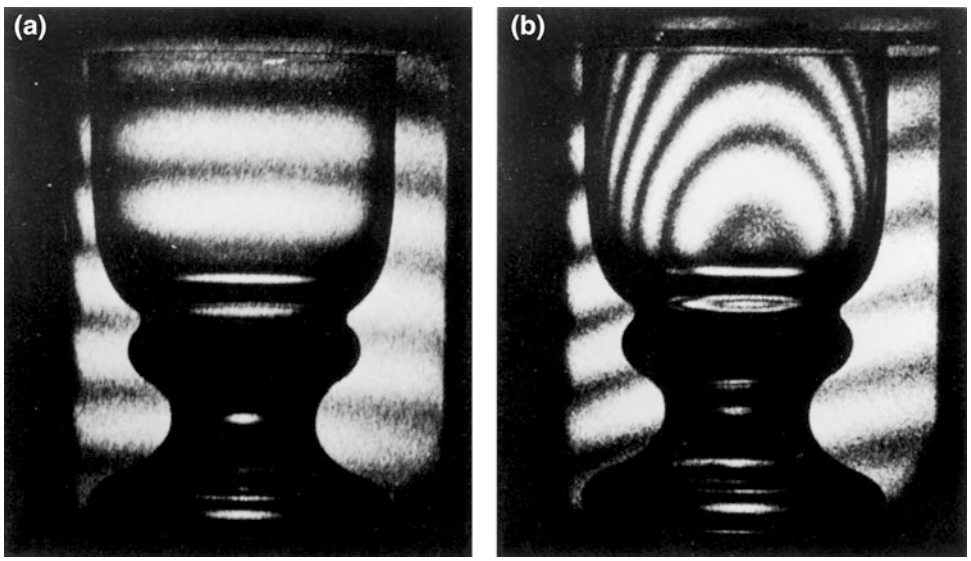


Fig. 12.21 Real time holographic interferometry **a**) interference between the original object wave and the wave reconstructed from the hologram **b**) superposition of two exposures of an empty glass and this glass filled with coal gas from a lighter

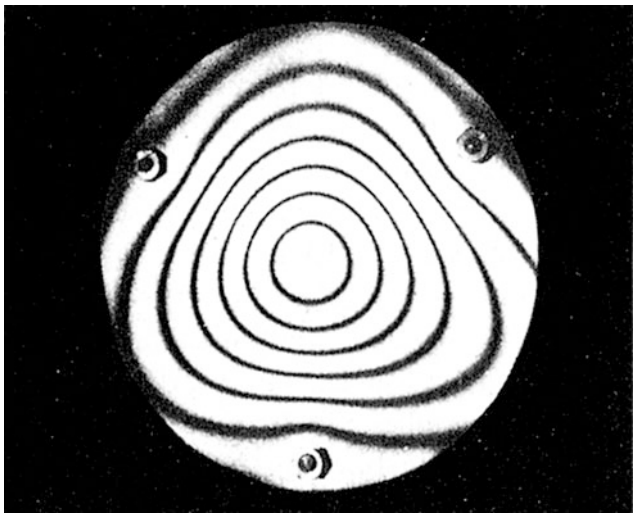


Fig. 12.22 Holographic interferogram showing the deformation of an aluminum disc. The hologram was exposed for 15 s before and after the deformation

phase difference of ½π the deformation is $\lambda/2$ and the double exposed hologram shows black lines of $\lambda/2$-deformations.

Another example is the double exposure hologram of a light bulb shown in Fig. 12.23 where the first hologram is taken when the light bulb is on and the second some seconds after it is switched out. The reconstructed hologram shows the convection of the filling gas above the filament and the thermal deformation of the glass bulb. The distance between two black lines corresponds to a deformation of $\lambda/2$.

(c) For periodically oscillating objects the exposure of the hologram is chosen longer than the oscillation period. Since the object stays longest in the turning points of the oscillation (because here the velocity of the oscillating parts is zero) the wave scattered by the object during this times contributes more to the exposure of the hologram than at other times. These positions of the object are therefore more distinct visible after the reconstruction of the hologram. The separation of the black lines corresponds to an oscillation amplitude of $\lambda/2$.

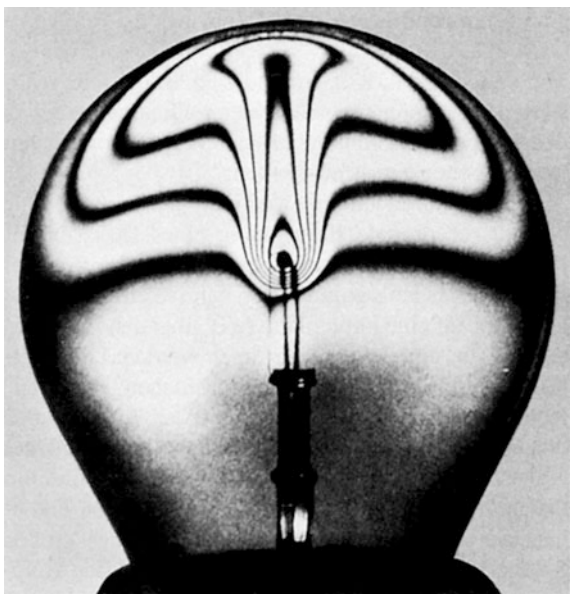


Fig. 12.23 Convection above the filament of a ligth bulb and thermal expansion of the heated bulb (from: M. Cagnet, M. Francon and S. Mallick: Atlas optischer Erscheinungen, Springer Berlin, Heidelberg 1971)

12.4.5 Applications of Holography

From the many possible and already realized applications of holography only a few will be discussed here besides those which have been already treated in the previous section.

An interesting application is the digital calculation of the hologram for ideal objects in their nominal condition. Such a digital hologram stored in a computer can then be transferred into a real hologram by printing it on a transparent foil. The superposition of this ideal hologram with that produced by exposure of the real object makes all deviations of the object from its nominal conditions visible. One example is the polishing procedure of large astronomical telescope mirrors. With this technique all locations on the mirror surface which have deviations from the ideal rotational paraboloid are simultaneously visible and can be removed by polishing. This shortens the polishing process considerably.

Applications in car industry are for instance double exposure holograms of car tires at different pressures, where small bulges due to locally changing tire wall thickness can be made visible.

With the holographic interferometry the growth of mushrooms can be measured within seconds when two holograms are taken within some seconds and superimposed on the photo plate. This allows the optimization of essential nutrient intake in fungal cultures.

An interesting application in medicine is the holographic survey of human skulls. Comparing such surveys with X-ray exposures the distribution of soft tissue can be deduced. Compiling a list of this distribution for different face shapes the surgeon of anomalies or injuries of the face can decide which affect the planned operation has onto the final natural appearance and he can optimize his operations accordingly [20].

In information technology the optimization of holographical storage media will become important. They have a larger information density because it is possible to superimpose many holograms in a small volume. A possible

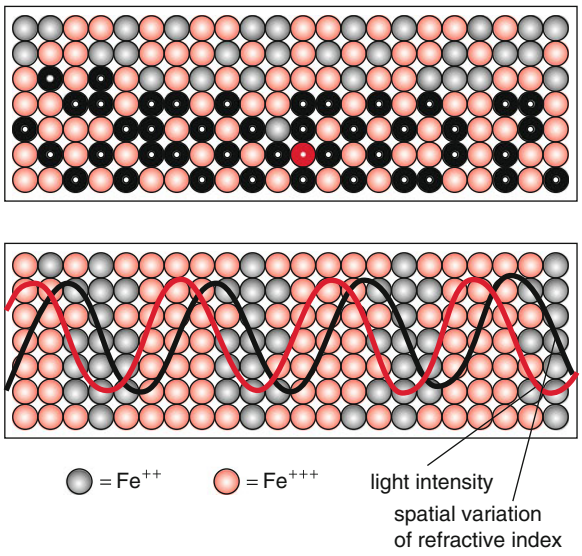


Fig. 12.24 Holographic storage in a Ferro-electric crystal $LiNBO_3$ doped with Fe^{2+} and Fe^{3+} ions **a**) before **b**) after exposure—The red curve gives the light intensity, the black curve the spatial distribution of the charge density, which influence the refractive index

storage material is for instance a Ferro-electric crystal. The electric field of the light wave constructing the hologram shifts the electric charges. The hologram does not appear as blackening pattern of a photo plate but as a space charge distribution which results in a spatial pattern of the refractive index (Fig. 12.24). Such holograms can be read out by a reference wave [21].

12.5 Fourier-Optics

Many problems in modern optics can be solved in an elegant mathematical way by Fourier-transformations. We have already seen in Sect. 10.8 that the spatial amplitude distribution of the Fraunhofer diffraction pattern in the observation plane can be regarded as the Fourier transform of the amplitude distribution of the transmitted incoming wave. We will show here that the imaging lens acts as a Fourier-lens, which transforms the object plane into the Fourier-plane where the diffraction image of the object is formed. When this Fourier-plane is further imaged by a second lens the original object is reproduced because the Fourier-transform of the Fourier-transform of a function f reproduces the original function f, but with reversed sign of the arguments.

$$\mathcal{F}[\mathcal{F}[f(x,y)]] = f(-x,-y).$$

The image is therefore inverted. The essential point is now, that the diffraction image can be altered by apertures, filters or phase plates. This leads to corresponding changes of the real image of the object. Such an optical filtering can, for instance, enhance the contrast of structures in the object image, or it can reduce incommoding background structures. This leads to an improvement of the image quality. Finer details, which could have been masked by background perturbations, can be made clearly visible by background subtraction.

For more details the reader is pointed to the literature [22, 23] ….

12.5.1 The Lens as Fourier-Imaging Component

We regard in Fig. 12.25 a plane wave with a wave vector $\boldsymbol{k}$. Its x-component forms the angle α and its y-component the angle β against the z-direction. The wave is imaged by the lens L into the plane $z = f_B$. According to (10.88) the amplitude distribution in this Fourier-plane

$$E(x',y') = A(x',y',f_B) \cdot \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} E(x,y)\mathrm{e}^{-2\pi\mathrm{i}(\nu_x x + \nu_y y)}\mathrm{d}x\mathrm{d}y \tag{12.10}$$

is given by the Fourier-transform of the field distribution $E(x, y)$ in the object plane, where the phase factor A is

$$A = \mathrm{e}^{\mathrm{i}kz}\mathrm{e}^{(\mathrm{i}\pi/\lambda z)\cdot(x'^2+y'^2)}$$

with $|A| = 1$.

The quantities

$$\nu_x = \frac{x'}{\lambda z} = f_B\frac{\tan\alpha}{\lambda z} \approx \frac{\alpha}{\lambda} \tag{12.11a}$$

$$\nu_y = \frac{y'}{\lambda z} = f_B\frac{\tan\beta}{\lambda z} \approx \frac{\beta}{\lambda} \tag{12.11b}$$

for the plane $z = f_B$ and with the approximation $\tan\alpha \approx \alpha$ and $\tan\beta \approx \beta$ are called the **spatial frequencies** of the diffraction pattern in the Fourier-plane.

If the object plane is the front focal plane of the lens $L\,(z_0 = -f_B)$, the factor

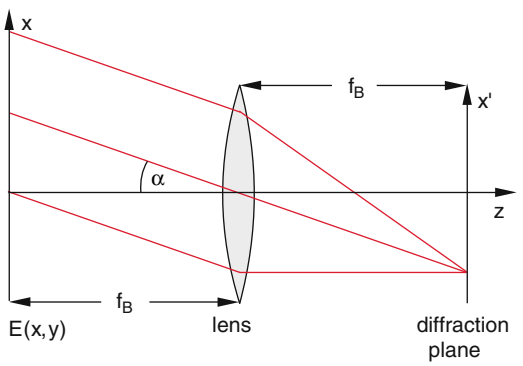


Fig. 12.25 The lens as Fourier-transformer of the field distribution $E(x, y)$ into the diffraction pattern $E(x', y') = E(\nu_x, \nu_y)$

$$A = e^{-ikf_B} e^{(i\pi f_B/\lambda)(\alpha^2+\beta^2)}$$

becomes independent of the location (x', y') in the Fourier plane. In this case the lens L performs, according to (12.10), exactly a Fourier transform of the front focal plane into the image focal plane.

The spatial intensity distribution in the observation plane

$$I(x', y') = |E(x', y')|^2 = |\mathcal{F}(E(x, y))|^2 \tag{12.12}$$

is the observed diffraction image of the object. Since for the intensity which is proportional to the absolute square of the amplitude, that phase factor becomes 1, any arbitrary plane in front of the lens can be chosen as the object plane (for instance directly before the lens).

The diffraction image is generally very small. Therefore one has to choose lenses with a large focal length f_B in order to generate an image with convenient size.

We will illustrate the Fourier-transform with a lens by some examples.

12.5.1.1 Point-like Light Source

A point-like light source at the point (x_0, y_0) in the front focal plane of the lens (Fig. 12.26) has the electric field amplitude distribution

$$E(x, y) = E_0 \delta(x - x_0)\delta(y - y_0), \tag{12.13}$$

where $\delta(x)$ is the delta function. Inserting into (12.10) yields the amplitude distribution in the image plane

$$\begin{aligned} E(x', y') &= A \cdot \int\int E_0 \delta(x - x_0)\delta(y - y_0) \\ &\quad \cdot e^{-2\pi i(v_x x + v_y y)} dx dy \\ &= A \cdot E_0 e^{-2\pi i(v_x x_0 + v_y y_0)}. \end{aligned} \tag{12.14}$$

This represents with

$$v_x = \frac{x'}{\lambda f_B} \approx \frac{\alpha}{\lambda};$$
$$v_y \approx \frac{\beta}{\lambda}$$

a plane wave with a wave vector $\boldsymbol{k}$, where the x-component and the y-component form the angles α and β against the z-axis.

The intensity

$$I \propto |E(x', y')|^2 = E_0^2$$

in the Fourier plane is constant, i.e. independent of x' and y'. The observation plane is uniformly illuminated.

12.5.1.2 Two Point-like Sources

We regard two point-like light sources in the object plane at the locations $(0, y_0)$ and $(0, -y_0)$ (Fig. 12.27). The field distribution in the object plane is then

$$E(x, y) = E_0 \delta(x)[\delta(y - y_0) + \delta(y + y_0)]. \tag{12.15}$$

The Fourier-transform in the image plane is then obtained from (12.10) analogous to (12.14) as

$$\begin{aligned} E(x', y') &= A \cdot (e^{-2\pi i v_y y_0} + e^{2\pi i v_y y_0}) \\ &= 2A \cdot \cos(2\pi v_y y_0) \end{aligned} \tag{12.16}$$

and the intensity distribution becomes

$$I(x', y') \propto 4A^2 \cos^2(2\pi v_y y_0) = 2A^2[1 + \cos(4\pi v_y y_0)]. \tag{12.17}$$

This represents a cosine grating with parallel stripes in the x-direction which have a space frequency

$$v_y = \frac{1}{2y_0} \tag{12.18a}$$

that is equal to the inverse distance of the two point light sources.

The spatial distance of the stripes in the focal plane of the lens

$$\Delta y' = v_y \cdot \lambda \cdot f_B \tag{12.18b}$$

is proportional to the focal length f_B of the imaging lens and to the wavelength λ.

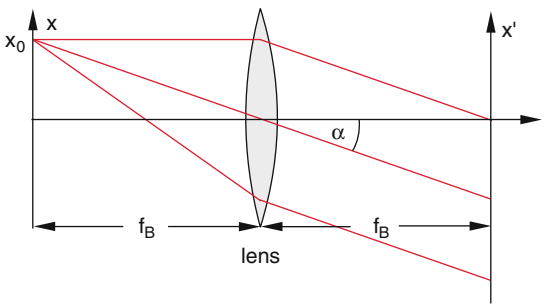


Fig. 12.26 Fourier-transformation of a point like source

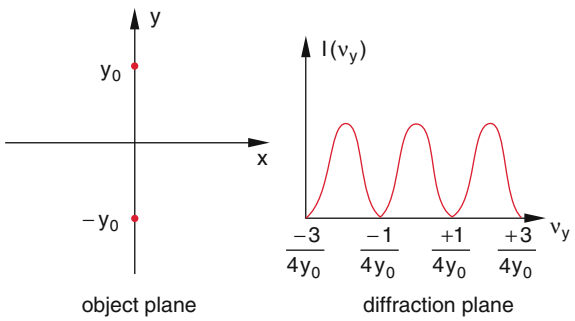


Fig. 12.27 Fourier-Frequency spectrum of two point sources

12.5.1.3 Optical Line Grating

The field amplitude of an incident plane wave in z-direction transmitted through a line grating with groove distance d is

$$E(x,y) = E_0 \cdot \sum_{n=1}^{N} \delta(x - nd) * \mathrm{rect}\frac{x}{a} \qquad (12.19)$$

where * means the convolution of the delta function with the step function rect (x/a). (rect$(x/a) = 1$ for $0 \le x/a \le 1$ and rect$(x/a) = 0$ elsewhere).

The amplitude distribution in the observation plane is obtained by inserting (12.19) into (12.10). This yields

$$E(x',y') = E_0 \cdot \delta(v_y) \cdot \frac{\sin \pi a v_x}{\pi v_x} \sum_{n=1}^{N} \mathrm{e}^{-2\pi i n \cdot d \cdot v_x}, \qquad (12.20)$$

This gives the intensity distribution, already known from Sect. (10.5.2)

$$I(v_x, v_y) \propto |E_0|^2 \delta v_y \cdot a^2 \cdot \frac{\sin^2(\pi a v_x)}{(\pi a v_x)^2} \cdot \frac{\sin^2(\pi N d v_x)}{\sin^2(\pi d v_x)} \qquad (12.21)$$

which is shown in Fig. 12.28. This illustrates that the coarse structure in the diffraction pattern i.e. the spatial frequency v_x which corresponds to the envelope of the interference maxima is caused by the narrow slit width a in the object plane whole the fine-structure, i.e. the high spatial frequency $N \cdot v_x$ which corresponds to the distance between the interference maxima is caused by the whole grating with the width $N \cdot d$, i.e. by a broad structure in the object plane.

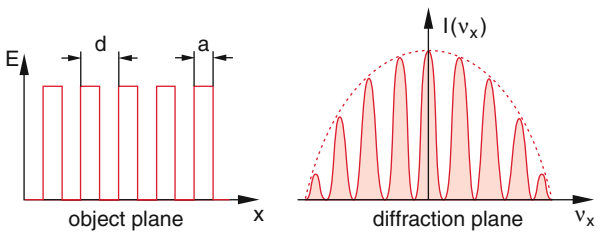


Fig. 12.28 Difraction pattern of a grating with rectangular long slits

Small spatial frequencies in the diffraction pattern are caused by broad spatial structures in the object plane, while fine structures of the object cause high spatial frequencies, i.e. large deviations in the diffraction plane.

12.5.2 Optical Filtering

The basic principle of optical filtering is illustrated by Fig. 12.29.

The lens L1 transforms the amplitude distribution in the object plane into a diffraction pattern in the focal plane of L1 which is equal to the Fourier-transform of the amplitude distribution $E(x, y)$. When the image plane of L1 is further imaged by a second lens L2 into the focal plane of L2 the image produced there is equal to the Fourier-transform of the Fourier-transform of the object plane. As has been shown earlier, this corresponds to the structure in the object plane. The two equal lenses with focal length $f_1 = f_2 = f$ and a distance $2f$ generate the real inverse image ($x \to -x$, $y \to -y$) of the object.

What is the difference of this procedure from a normal imaging of an object at the distance $2f$ by a single lens?

The essential point is that for the imaging in Fig. 12.29 a Fourier plane between the two lenses exists, where optical filters or aperture can be inserted which change the diffraction pattern in this plane, and which in turn modify the real image of the object in a characteristic but wanted way. This will be illustrated by some examples.

12.5.2.1 Low Pass Filter

We have seen in the previous section that fine structure details in the object plane result in high spatial frequencies in the diffraction plane. If these high spatial frequencies (which corresponds to large spatial deviations in the Fourier plane) are suppressed by an aperture, the often unwanted fine structures of the object do not appear in the real image. An example is an area, uniformly illuminated by the laser beam enlarged by the two lenses L_1 and L_2 (Fig. 12.30). The magnification of the diameter is given by the ratio f_2/f_1 of the focal lengths. Impurities (dust particles, streaks) or irregularities on the lens surfaces cause diffraction of the light into higher diffraction orders. The imaging by the second lens lead to a granular structure of the image which is often superimposed by diffraction rings of the dust particles. A pinhole in the focal plane of L_1 suppresses all higher diffraction orders and generates a uniform brightness of the enlarged laser beam. All filters which eliminate higher diffraction orders are called **low pass filters** following the

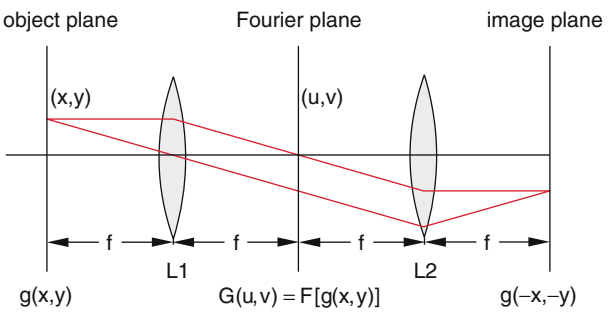


Fig. 12.29 Schematic representation of the optical Fourier-transformation by L_1 and the back transformation by L_2

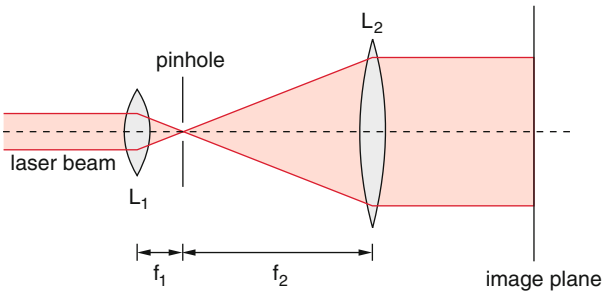


Fig. 12.30 Pinhole as low-pass space frequency filter to improve the quality of the enlarged laser beam

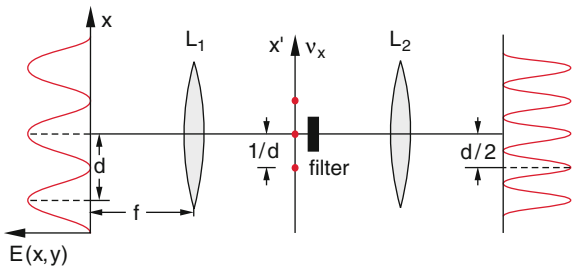


Fig. 12.31 High frequency pass filter at the imaging of a cosine grating

nomenclature in Electro-technics, where low pass filters suppress all higher frequencies. (see Sect. 5.5). The pinhole acts like a point-like light source in the focal plane of L_1, which generates a plane wave behind the lens L_2 and therefore a uniform intensity in the image plane.

12.5.2.2 High Pass Filter

High pass filters in the diffraction plane of the optical system suppress the lower diffraction orders, which are less deflected, but transmit the higher orders. This shall be illustrated by the imaging of a one-dimensional cosine grating with the grating period d in the x-direction. The electric field distribution in the object plane is

$$E(x) = E_0 \cos^2\left(\frac{\pi x}{d}\right) = \frac{E_0}{2}\left[1 + \cos\left(\frac{2\pi x}{d}\right)\right] \tag{12.22a}$$

This can be written in the form

$$E(x) = E_0\left[\frac{1}{2} + \frac{1}{4}e^{2\pi i x/d} + \frac{1}{4}e^{-2\pi i x/d}\right] \tag{12.22b}$$

Inserting this into (12.10) one obtains after a short calculation for the Fourier-transform, i.e. the amplitude distribution in the diffraction plane

$$E(v_x, v_y) = \frac{E_0}{2}\delta(v_y)\left[\delta(v_x) + \frac{1}{2}\delta\left(v_x - \frac{1}{d}\right) + \frac{1}{2}\delta\left(v_x + \frac{1}{d}\right)\right], \tag{12.23}$$

where the three terms represent the 0th, the +1st and the-1st diffraction orders. The diffraction pattern consists of 3 points on the v_x-line with spatial frequencies $v_x = 0$ and $v_x = \pm 1/d$. In the diffraction plane these points are located at the points $x' = 0$ and $x' = \pm f \cdot \lambda/d$ (red points in Fig. 12.31). The lens L_2 images this diffraction pattern into the image of the original cosine grating.

If now the 0th diffraction order is suppressed by a small opaque disc the first term in (12.23) is missing. The Fourier-transform of this new diffraction pattern

$$E(v_x, v_y) = \frac{E_0}{4}\delta(v_y)\left[\delta\left(v_x - \frac{1}{d}\right) + \delta\left(v_x + \frac{1}{d}\right)\right] \tag{12.24}$$

with $v_x = x'/(\lambda \cdot f)$, $v_y = y'/(\lambda \cdot f)$ is obtained by inserting this into (12.10). This gives the field distribution in the image plane

$$E(x) = \frac{1}{2}\cos\left(\frac{2\pi x}{d}\right) \Rightarrow I(x) = \frac{1}{4}\cos^2\left(\frac{2\pi x}{d}\right). \tag{12.25}$$

This is again a cosine grating, which has, however, only the half period. There are twice as many maxima and minima as for the field distribution (12.23).

This high pass filtering can be used also for making transparent objects visible. Although there is no essential attenuation of the transmitted light but a phase shift does occur. If such objects are observed without filtering the phase factor for the intensity $I \propto |E|^2$ becomes 1 and the object is not noticeable.

This is no longer the case, if for example the 0th diffraction order is suppressed. This can be understood as follows:

For a transmission

$$\tau(x, <y) = a \cdot e^{i\varphi(x,y)}, \tag{12.26}$$

of the object wave the spatial amplitude distribution in the object plane becomes

$$E(x, y) = E_0 \cdot \tau(x, y) = a \cdot E_0 e^{i\varphi(x,y)} \tag{12.27}$$

The phase factor $\varphi(x, y)$ contains the information about the object. The transmitted intensity

$$I(x, y) \propto |E|^2 = a^2 \cdot E_0^2 = I_0$$

does not depend on x or y i.e. it is constant over the whole area of the image.

For small phase shifts ($\varphi(x,y) \ll 1$) we can expand the exponential function in (12.27) and obtain the approximation

$$E(x,y) = a \cdot E_0[1 + \mathrm{i}\varphi(x,y)], \tag{12.28}$$

The Fourier-transform of (12.28) gives the amplitude distribution in the diffraction plane

$$\begin{aligned} E(v_x, v_y) &= \mathcal{F}[E(x,y)] \\ &= a \cdot E_0\{\delta(v_x)\delta(v_y) + \mathrm{i}\mathcal{F}[\varphi(x,y)]\} \end{aligned} \tag{12.29}$$

If the 0th diffraction order is suppressed, the first term in (12.29) becomes zero. The amplitude distribution in the image plane after imaging the diffraction plane by the lens L_2 into the image plane becomes

$$\begin{aligned} E(x_B, y_B) &= \mathrm{i}E_0 \cdot a \cdot \mathcal{F}[\mathcal{F}[\varphi(x,y)]] \\ &= \mathrm{i}a \cdot E_0\varphi(-x,-y) \end{aligned} \tag{12.30}$$

and the intensity distribution in the image plane is

$$I(x_B, y_B) \propto |E(x_B, y_B)|^2 = a^2E_0^2\varphi^2(-x,-y), \tag{12.31}$$

This shows that the phase information is saved and the object becomes visible.

12.5.3 Optical Pattern Recognition

Optical filtering is also very useful for optical pattern recognition . One possible application is the inspection of produced pieces (e.g. small cogwheels or punched forms) in a serial inspection of a large number of pieces [24, 25].

It must be assured that for each piece the technical tolerances are kept within the demanded limits. Every piece must be compared with a model that has been precisely produced and shows no deviations from the desired values.

At first a hologram of this model is taken by placing the hologram plate in the diffraction plane of the model as object and illuminating it with a plane reference wave. Such a hologram is called *Fourier-hologram* and corresponds to the Fourier-transform of the model. The developed hologram is now placed in the filter plane (i.e. the diffraction plane of the pieces to be inspected). In the image plane the superposition of the images of the model and the real object can be seen. Choosing the correct phase one can reach that the difference of the two images appears, which immediately shows the deviation of the inspected piece from the model.

An important biological application of this pattern recognition is the fast distinction between healthy and cancer cells, which is very useful for serial histological examinations for early cancer diagnosis.

If objects change in time (for instance deformation of work pieces under external pressure, or the change of the cloud structure in Jupiter's atmosphere) this can be readily inspected by such methods of pattern recognition. All images of objects at a later time are compared with the image at time t_0 in a similar way as described before. Only changes appear in the difference image while all constant forms are suppressed.

A drawback of this method is its extreme sensitivity against even very small shifts of the filter which can alter the pattern which should be recognized.

12.6 Micro-Optics

Micro-Optics is a modern branch of optics which started around 1980 and has meanwhile undergo a very rapid development reaching technical maturity. This dynamic expansion was only possible through the parallel proceeding of micro-technology (lithographic techniques, mechanical production of micro-structures) and due to a better understanding of the physical phenomena which occur at the optical imaging and wave propagation through micro-optical elements.

In this section we will only briefly deal with some aspects and applications of micro-optics [26, 27].

12.6.1 Diffractive Optics

In Chap. 9 we have discussed that light can be deflected by prisms or collected by lenses. Both effects are based on the wavelength dependent light refraction at interfaces between two media with different refractive index. All imaging techniques which are based on refraction are called *refractive optics.*

During the last years new fabrication techniques of micro-mechanics and microelectronics together with new methods of computer aided design have facilitated the creation of new optical elements which are based on diffraction instead of refraction. They can perform deviation as well as focusing of light beams. Such imaging methods based on microscopic small optical elements are called **diffractive optics**. Its principle will be illustrated by some examples.

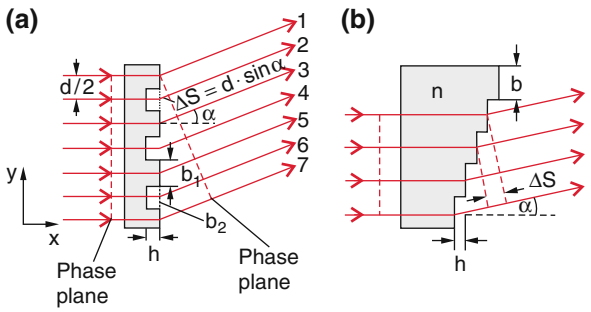


Fig. 12.32 **a**) Diffraction of light at a strip-type phase plate where b_2 is the width and h the depth of the grooves, b_1 the width of the stripes and d the distance between two grooves. **b**) Diffraction by a graded profile

In Fig. 12.32a a glass plate is shown, where stripes of rectangular grooves with width b and depth h are etched into the glass surface. When a plane wave is incident onto the glass plate, an optical path difference $\Delta s = (n-1)h$ appears between vertically incident light passing through the grooves with depth h and light that passes through the stripes between the grooves. The glass plate therefore acts as a phase grating. While for a transmission grating with equal slits and strips the transmitted amplitude is spatially modulated in y-direction and the transmission is $T = 0.5$, for the phase grating the phase is spatially modulated in y-direction and the transmission is $T = 1.0$. Analog to the classical optical grating each stripe with width b acts with respect to diffraction as a slit where the central diffraction order is deflected within the angular range $|\sin\alpha| \leq \lambda/b$. If sufficiently large deflection angles are demanded, the stripe width b must be of the same order as the wavelength λ.

The partial waves from the different stripes add to a macroscopic wave if the phase difference between the waves from adjacent stripes is $\Delta\varphi = m \cdot 2\pi$, which implies that the optical path difference is $\Delta s = m \cdot \lambda$. From Fig. 12.32a we can see, that the path difference between adjacent beams for $m = 2, 4, 6, \ldots$ = even and between adjacent odd beams ($m = 1, 3, 5, \ldots$) in the direction α is $\Delta s = d \cdot \sin\alpha$ with $d = b_1 + b_2$. For constructive interference we therefore get the condition.

$$d \cdot \sin\alpha = \pm m_1 \cdot \lambda \qquad (m_1 = 0, 1, 2, \ldots). \tag{12.32a}$$

The odd-numbered beams have the optical path difference

$$\Delta s_2 = \pm(n-1) \cdot h \pm \frac{1}{2} d \sin\alpha \tag{12.32b}$$

against the adjacent even numbered beams.

All odd-numbered partial beams can interfere constructively with the even numbered beams if $\Delta s_2 = m_2 \cdot \lambda$, $(m_2 = 0, 1, 2)$. The subtraction of (12.32b) and (12.32a) gives

$$\pm(n-1) \cdot h + \frac{1}{2} d \sin\alpha = (m_2 - m_1) \cdot \lambda. \tag{12.33}$$

For $m_2 - m_1 = 0$ the direction for the zeroth order constructive interference becomes

$$\sin\alpha_0 = \pm \frac{2(n-1)h}{d}. \tag{12.34a}$$

Here the optical path difference is zero (besides dispersion effects) and $\sin\alpha_0$ is independent of λ. For the interference maximum of 1. Order ($m_2 - m_1 = \pm$) the deflection angle is

$$\sin\alpha_1 = \frac{(n-1)h \mp \lambda}{d/2}. \tag{12.34b}$$

Examples

$n = 1.5$, $h = 1.5\,\mu\text{m}$, $b_1 = b_2 = 1\,\mu\text{m} \Rightarrow d = 2\,\mu\text{m}$; $m_2 - m_1 = 0 \Rightarrow \sin\alpha_0 = 0.75 \Rightarrow \alpha_0 = \pm 48.6°$ independent of λ.

(a) For a wavelength $\lambda = 0.5\,\mu\text{m}$ we obtain for $m_2 - m_1 = 1$: $\sin\alpha_1 = 0.25 \Rightarrow \alpha_1 = 14.5°$. For $m_2 - m_1 = -1$ we get $\sin\alpha_1 = 1.25$ which shows that there is no interference maximum for the-1st order.

(b) With a stripe heights $h = 1\,\mu\text{m}$ one obtains:

For $m_2 - m_1 = 0 \Rightarrow \sin\alpha_0 = 0 \Rightarrow \alpha_0 = 0$
For $m_2 - m_1 = +1 \Rightarrow \alpha_1 = 0$
For $m_2 - m_1 = -1 \Rightarrow \alpha_1 = 90°$.
For this example there is only a constructive interference in the forward direction $\alpha = 0$ for the zeroth—as well as for the first order.

This example illustrates that the deflection of the transmitted light against the incident light can be chosen within wide angular ranges

$$\alpha_0 = \arcsin\left(\frac{2(n-1)h}{d}\right), \tag{12.34c}$$

by selecting the proper values of grating constant d and stripe heights h. Since the refractive index n depends slightly on λ there remains a weak dependence of α_0 from λ.

The transparent plate can be also constructed with a stepped profile as shown in Fig. 12.32b. The phase plane of the transmitted wave is inclined by the angle α against that of the incident wave and is determined by the condition that the optical path difference $\Delta s_1 = n \cdot h$ is just compensated by the corresponding path difference $\Delta s = b \cdot \sin\alpha$ behind the plate. The zeroth interference order therefore occurs for

$$n \cdot h - b \cdot \sin\alpha_0 = 0 \Rightarrow \sin\alpha_0 = n \cdot h/b. \tag{12.35}$$

Example

$n = 1.5$, $h = 0.2\,\mu\text{m}$, $b = 1\,\mu\text{m} \Rightarrow \sin\alpha_0 = 0.30 \Rightarrow \alpha_0 = 17.5°$.

Note that also for the zeroth interference order the deflection angle α_0 depends slightly on λ because of the dispersion $n(\lambda)$. This dependence is, however, for the zeroth order much smaller than for higher orders. For the mth−interference order is

$$\begin{aligned} &n \cdot h - b \cdot \sin\alpha = \pm m \cdot \lambda \\ &\Rightarrow \sin\alpha = (n \cdot h \mp m \cdot \lambda)/b. \end{aligned} \tag{12.36}$$

For the example above the interference angles for the first two diffraction orders appear for $\lambda = 0.5$ µm at

$$\alpha_1(m = +1) = -11.5^\circ; \alpha_2(m = -1) = +53^\circ.$$

Parallel incident light falling onto such a stepped profile plate is split into several partial beams with deflection angles determined by (12.36). The number of these partial beams depends on the width b of the steps. The diffraction at each single step limits the angular range, where only deflected light can be observed for $\sin\alpha \leq q\lambda/b$.

Example

For $b = 1\,\mu m$ and $\lambda = 0.5\,\mu m$ the angular range is limited to $|\alpha| < 30^\circ$. This means that the total transmitted intensity is distributed among the 0th and the 1st interference order. In this case only the 0th and the 1st order appear. The incident beam is split into two transmitted partial beams. For smaller values of h and b higher orders are realized and more deflected beams are obtained.

12.6.2 Fresnel Lenses and Lens Arrays

As a second example we will discuss a Fresnel lens, which acts as a Fresnel zone plate (Sect. 10.6.2).

For Fresnel zones the path difference between the observation point and two adjacent zones is $\lambda/2$. They can be realized by etching circular grooves into the surface of a circular transparent glass plate (Fig. 12.33). The difference $\Delta s_m = s_{m+1} - s_m$ for the path length from adjacent grooves to the focal point F can be expressed by the radii of the grooves

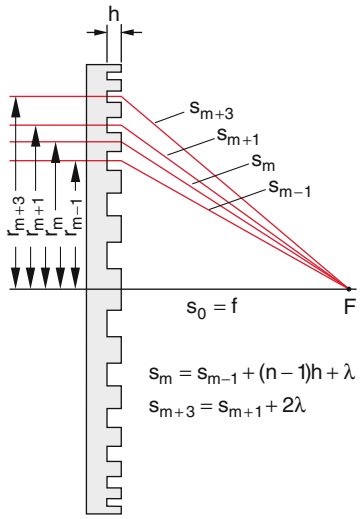


Fig. 12.33 Glass plate with circular grooves. Acting as Fresnel lens

$$r_m^2 = s_m^2 - s_0^2 = (s_0 + m \cdot \lambda/2)^2 - s_0^2 \approx s_0 \cdot m \cdot \lambda \quad (12.37a)$$

which is valid for $s_0 \gg m \cdot \lambda$. We can therefore write for the radius of the mth-Fresnel zone (as already derived in (10.64))

$$r_m = \sqrt{m \cdot s_0 \cdot \lambda}. \quad (12.37b)$$

The area A of each zone

$$A = \pi\left(r_{m+1}^2 - r_m^2\right) = \pi \cdot s_0 \cdot \lambda \quad (12.38)$$

is independent of m and therefore equal for each zone..

The path difference between a strip zone and a groove zone is

$$\Delta s = (n-1)h - \lambda/2.$$

If the heights h of the stripe is chosen such that $(n-1) \cdot h = \lambda/2$ then all partial waves from stripes and grooves are in phase at the point F and interfere constructively. Such a zone plate then acts as a lens with the focal length

$$f = s_0 = r_1^2/\lambda, \quad (12.39)$$

Note that the focal length depends on the wavelength λ and the radius r_1 of the first Fresnel zone.

Remark For the Fresnel-zone plate, discussed in Sect. 10.6.2 with alternate transparent and opaque zones the zones producing destructive interference had to be masked. This causes a 50% decrease of the transmitted intensity. For the phase modulated zone plate, however, the destructive interference is converted into a constructive one just by choosing the correct phase shifts. Therefore the total incident intensity is transmitted, which implies a gain by a factor 2 compared to the amplitude modulated zone plate.

From Eq. (12.39) we learn that the focal length f of *diffractive* lenses decreases with increasing wavelength λ contrary to *refractive* lenses where in the range of normal dispersion the refractive index decreases with increasing wavelength and therefore the focal length increases. By an appropriate combination of the diffractive and refractive effect of a lens, an achromatic lens can be realized (Fig. 12.34).

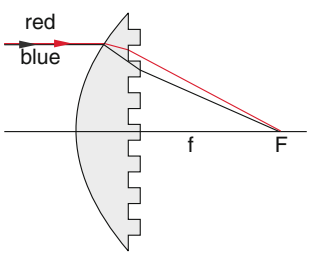


Fig. 12.34 Combination of a diffractive Fresnel lens and a refractive classical lens for the realization of an achromatic lens

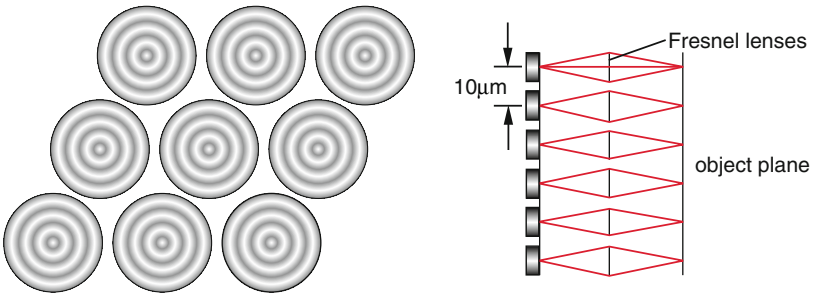


Fig. 12.35 Two-dimensional array of small Fresnel lenses for the simultaneous imaging of large object areas without aberrations onto a photodiode array in the image plane

Besides their possible small sizes a great advantage of diffractive lenses is the possibility of cheap mass production based on etching techniques which have been already optimized in micro-electronics for the production of microchips. The calculated optimization of the surface structure can minimize imaging optical aberrations. Diameter and focal length of such micro lenses can be made quite small (e.g. below 1 mm). They can be therefore used in medicine as imaging elements in endoscopes (these are optical fiber systems that can be inserted into the human body for medical inspection and treatment).

A whole array of Fresnel lenses can be arranged on a glass plate (Fig. 12.35). This allows the simultaneous imaging of many different sections of an extended object. If a photodiode array is placed in the image plane, each diode receives the signal of the corresponding section of the object. Since Fresnel lens arrays as well as photodiode arrays can be produced with integrated techniques, imaging system and detector unit can be produced on a single chip, which reduces the production costs considerably [28].

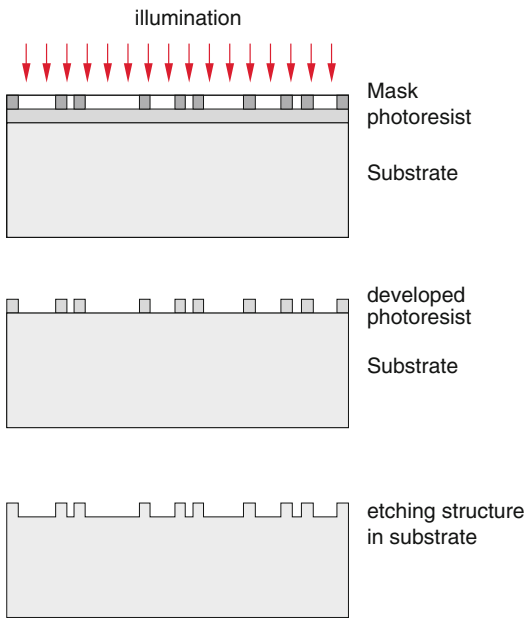


Fig. 12.36 Lithographic technique for the production of micro-optics

12.6.3 Production Techniques of Diffractive Optical Elements

The most important technique for the production of diffractive optical elements is the lithography. At first a heavily scaled down copy of the original pattern is produced on a photosensitive layer by illumination with ultraviolet light. The illuminated parts appear black after the development of the photo layer. This structured layer is now used as mask which is placed on top of a thin glass plate with a photo-layer on its surface (Fig. 12.36). The photo-layer is now illuminated through the mask and the illuminated parts are removed by etching reagents. These parts are now accessible to further treatment by other etching substances which produce the wanted pattern (e.g. the circular zones of a Fresnel lens) on the glass plate.

Besides by this photographic technique the mask can be also produced by an electron beam which is scanned across the mask plane and generates there the wanted pattern.

12.6.4 Refractive Micro-Optics

Besides lenses based on diffractive optics , micro lenses can be also produced which use the refraction for the imaging process, just as for macro lenses. One can distinguish between two different classes of micro lenses:

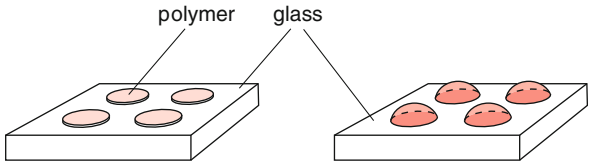


Fig. 12.37 Fabrication of micro-lenses of plastic by heating and melting of micro-cylinders which have been produced by lithographic techniques on a glass substrate

(a) lenses with a given surface structure which perform imaging of light beams
(b) lenses with a radial gradient of the refractive index which are made of an inhomogeneous material.

Refractive micro-lenses can be for instance produced by a lithographic process (Fig. 12.37). Plastic micro-cylinders with diameter D and heights h are placed on a glass surface. If they are heated above their melting temperature the cylinders are transformed into the liquid phase. Due to surface tension the liquid takes up a form with minimum surface for a given volume. This is a spherical cap with a radius of curvature, which depends on the quantities D and h and which acts as a micro-lens on the glass surface. By selecting the dimensions D and h of the cylinder one can define the radius of curvature R and therefore the focal length

$$f = R/(n-1)$$

of the micro-lens.

Using cylinders with a gradient of the refractive index, which can be produced by diffusion of impurity atoms into the plastic material, they can be used as cylindrical lenses.

The diameter of the micro lenses range between 5 and 500 µm, their focal length is between 50 µm and some millimeters [29].

12.7 Optical Waveguides and Integrated Optics

The realization of integrated optics takes advantage of the already existing micro structures (in the micrometer to nanometer range) of integrated electronic devices on the surface or in the interior of a substrate. Here light is coupled into and guided through tiny waveguides in the µm-range, where it is modulated or transferred into adjacent wave-guides. By this means an arbitrary optical input signal can be structured, encoded or transformed into any wanted form. It can be also distributed into many exit channels.

We will now discuss this in more detail:

12.7.1 Light Propagation in Optical Waveguides

Essential technical principles of integrated optics are based on the propagation of optical waves in waveguides, which has been already discussed in Sect. 7.9. Here are, however, no hollow waveguides with electrical conducting walls are used as shown in Fig. 7.26, but stripes or rectangular channels made of transparent, i.e. non conducting material with a refractive index, which is larger than that of the surroundings. The optical wave is therefore enclosed in and guided through the waveguide by total reflection [30].

All waves propagating through a waveguide with refractive index n must obey the wave equation

$$\Delta \boldsymbol{E} = \frac{n^2}{c^2}\frac{\partial^2 \boldsymbol{E}}{\partial t^2} \tag{12.40}$$

where Δ is the Laplace operator. The selection of special solutions is determined by boundary conditions, which depend on the dimensions and the refractive index of the waveguide and its surroundings.

As example we regard a planar waveguide between the planes $x = 0$ and $x = a$, which extends in the y-direction from $y = -\infty$ to $y = +\infty$ with the refractive index n_2 of the waveguide between two surrounding media with refractive indices n_1 and n_3 (Fig. 12.38). For a TE-wave (see Sect. 7.9) with a wave vector $\boldsymbol{k}$ forming the angle ϑ against the z-direction we get the expression

$$E(x,y,z,t) = E_y(x)\mathrm{e}^{\mathrm{i}(\omega t - \beta z)} \tag{12.41}$$

with the propagation constant $\beta = k \cdot n_2 \cos\theta$ and the wavenumber $k = \omega/c$. The amplitude $E_y(x)$ which depends solely on x, is given in the three ranges by

$$E_y(x) = \begin{cases} A \cdot \mathrm{e}^{px} & \text{for } x \le 0, \\ B \cdot \cos(hx) + C \cdot \sin(hx) & \text{for } 0 \le x \le a, \\ D \cdot \mathrm{e}^{-q(x-a)} & \text{for } x \ge \mathrm{a} \end{cases} \tag{12.42}$$

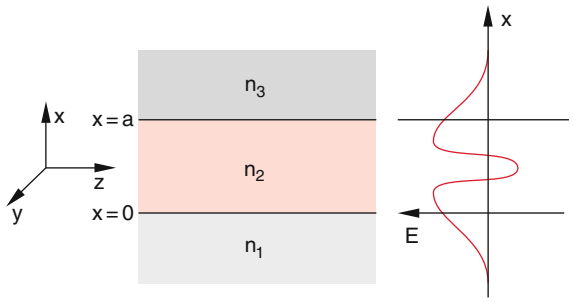


Fig. 12.38 Planar wave guide made as a layer with refractive index n_2 between layers with n_1 and n_3

This can be proved by inserting (12.42) into (12.40).

The continuity condition for $E_y(x)$ at $x = 0$ and $x = a$ demands $B = A$ and $D = B \cdot \cos(ha) + C \cdot \sin(ha)$. The additional condition, that also $\partial E_y/\partial x$ should be continuous, gives for $x = 0$ the relation $A \cdot p = C \cdot h$ where p and q are real numbers. We can therefore express all amplitudes by A and obtain:

$$E_y(x) = \begin{cases} A \cdot e^{px} & \text{for } x \le 0, \\ A \cdot [\cos(hx) + \frac{p}{h}\sin(hx)] & \text{for } 0 \le x \le a, \\ A \cdot [\cos(ha) + \frac{p}{h}\sin(ha)]e^{-q(x-a)} & \text{for } x \ge a. \end{cases} \tag{12.43}$$

Inserting (12.43) into the wave Eq. (12.40) gives

$$\begin{aligned} p &= \sqrt{\beta^2 - n_1^2 k^2}, \\ h &= \sqrt{n_2^2 k^2 - \beta^2}, \\ q &= \sqrt{\beta^2 - n_3^2 k^2}. \end{aligned} \tag{12.44}$$

All three quantities p, h and q are therefore determined solely by the wavenumber k, the refractive indices n_i and the parameter $\beta = k \cdot n_2 \cdot \cos\theta$ which gives the propagation parameter in the direction ϑ.

From the continuity of $\partial E_y/\partial x$ at $x = a$ we get the additional condition

$$\begin{aligned} &-h\sin(ha) - (q/h)\cos(ha) \\ &= -p[\cos(ha) + (q/h)\sin(ha)], \end{aligned} \tag{12.45a}$$

which gives

$$\tan(ha) = -\frac{p+q}{h(1 - qp/h^2)} \tag{12.45b}$$

Together with (12.43) this sets up a relation between the quantities p, q and h and illustrates that not arbitrary values of the propagation parameter β are allowed but only discrete values β_n which describe the allowed propagation modes of a wave in the waveguide.

From (12.42) we can see, that the wave can penetrate into the adjacent regions on both sides of the waveguide. Its amplitude decreases, however, exponentially.

The question is now, which propagation modes with wavelength λ can be guided through the planar waveguide with width a, without walking out of the waveguide? This depends on the difference $\Delta n = n_2 - n_1$ resp. $n_2 - n_3$ between the refractive indices of waveguide and surrounding material.

For practical applications symmetric waveguides with $n_1 = n_3 = n$ are standard. This implies according to (12.44) $\beta = n_1 \cdot k = n_3 \cdot k = n \cdot k$. With (12.45) we obtain the boundary condition for symmetrical planar waveguides

$$\tan(ha) = 0 \Rightarrow ha = m_s \cdot \pi. \tag{12.46}$$

For the coefficients h and a (12.46) gives the relations

$$\begin{aligned} h &= \sqrt{n_2^2 k^2 - \beta^2} = k\sqrt{n_2^2 - n^2} \\ &= \frac{2\pi}{\lambda}\sqrt{n_2^2 - n^2} \end{aligned} \tag{12.47}$$

$$a = (m_s \cdot \lambda/2) \cdot \sqrt{(n_2^2 - n^2)} \tag{12.48}$$

The integer m_s gives essentially the ratio of width a of the waveguide to the half wavelength $\lambda/2$.

Inserting (12.31) into (12.46) we obtain with $n_2^2 - n^2 = (n_2 - n) \cdot (n_2 + n)$ the minimum difference of the refractive indices

$$\Delta n = n_2 - n > \frac{m_s^2 \lambda^2}{4a^2(n_2 + n)}, \tag{12.49}$$

that is necessary to keep the wave mode with the mode parameter m_s inside the symmetric waveguide. This relation allows the calculation of

(a) the minimum difference Δn for a given wavelength λ and mode parameter m_s
(b) the maximum number m_s of modes that can be still guided for given values of Δn and wavelength λ.

Examples

1. $m_s = 0 \Rightarrow h = 0 \Rightarrow \beta = n_2 k \Rightarrow \vartheta = 0$. The wave vector points into the z-direction and the wave propagates parallel to the walls of the waveguide. There is no wavelength limit, but with increasing values of λ/a the fraction of the wave, travelling outside the waveguide, increases (Fig. 12.39).

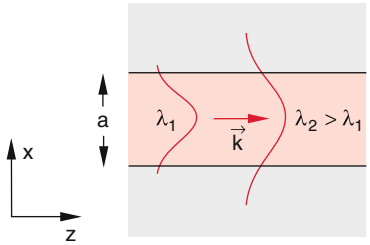


Fig. 12.39 Amplitude distribution $E(x)$ of the lowest TE-mode with $m_s = 0$ for two different wavelengths $\lambda_1 < a/2$ and $\lambda_2 > a/2$

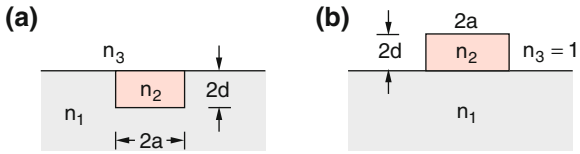


Fig. 12.40 Two possible realizations of dielectric stripe conductors **a**) imbedded **b**) superimposed

2. $m_s = 1, \quad \lambda = 600\,\text{nm}, \quad a = 2\,\mu m, \quad n = 1.5 \Rightarrow \Delta n > 0.0075 \Rightarrow n_2 = 1.5075$. This illustrates that already a very small difference Δn is sufficient to guide a wave mode with $m_s = 1$. From (12.44) we obtain for $\vartheta = 0$ the relations

$$p = \sqrt{\beta^2 - n^2k^2} = k\sqrt{n_2^2 - n^2} = \frac{2\pi}{\lambda}\sqrt{0.0215} \approx \frac{0.9}{\lambda}$$

$$\Rightarrow E_y(x<0) = A \cdot e^{-0.9x\lambda}$$

$$\Rightarrow I \propto |E_y|^2 = A \cdot e^{-1.8x/\lambda}.$$

The penetration depth into the surrounding material is about $\lambda/2$ which means that for $x = \lambda/2$ the amplitude has decreased to 1/e of its value inside the waveguide.

If the wave penetrates into the surrounding, its phase velocity is not only determined by the refractive index n_2 of the waveguide, but also by the index n of the surrounding material. Taking this into account the propagation parameter can be written as

$$\beta = k \cdot n_{\text{eff}} \cdot \cos\vartheta$$

Planar waveguides, where the wave is only limited in one direction, are a special class of more general waveguides with rectangular cross section where the wave is limited in x- and in y-direction. Possible technical realization are shown in Fig. (12.40). For instance, a groove with width $2a$ and depth $2d$ can be etched into a bulk material with refractive index n_1. Above the design either air with $(n_3 = 1)$ is the third region or the system is coated with another material with $n_3 > 1$. it is also possible to deposit a stripe of material with refractive index n_2 onto the planar surface of the bulk material, where the lower border has the refractive index n_1 and the sides the index $n = 1$ of air.

12.7.2 Modulation of Light

When a dielectric material is brought into an external electric field, its refractive index n changes. This gives the basis for

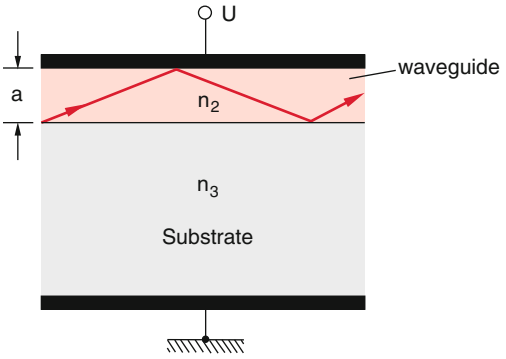


Fig. 12.41 Electro-optical modulation of a wave guide

electro-optical modulation of light in wave guides. Assume a waveguide with width a and refractive index n_2 is applied to a bulk material with refractive index n_3. When a voltage U is applied between two electrodes (Fig. 12.41) the refractive index of the material between the electrodes changes. If the bulk material has a non-vanishing electric conductivity the whole voltage app ears across the nonconductive wave guide and the electric field strength in the waveguide is $|\mathbf{E}| = U/a$.

As can be shown [40] the change of the extra-ordinary refractive index is

$$\Delta n_{\text{EO}} = n_2^3 \cdot \alpha_{\text{E}} \cdot U/a, \tag{12.50}$$

where α_{E} is a factor that depends on the polarizability of the material in the waveguide.

If the refractive indices n_2 and n_3 are chosen such that without electric filed the condition (12.49) is fulfilled for $m_s = 0$ but not for $m_s = 1$ no wave mode with $m_s = 1$ can propagate through the waveguide. Increasing the difference

$$\Delta n = n_2 + \Delta n_{\text{EO}} - n_3 \tag{12.51}$$

the condition (12.49) can be also satisfied for $m_s = 1$. This means that the wave mode with $m_s = 1$ can be switched on and off by the electric field.

12.7.3 Coupling Between Adjacent Waveguides

If two waveguides are separated by a thin layer with refractive index n_2 a wave which has been coupled into waveguide 1 and propagates there into the z-direction can extend with exponentially decreasing amplitude into waveguide 2. Part of the wave energy is then transferred to waveguide 2. We obtain for the electric field amplitudes

$$\frac{dA_1(z)}{dz} = -i\beta_1 A_1 + i\kappa_{12} A_2(z), \tag{12.52a}$$

$$\frac{dA_2(z)}{dz} = -i\beta_2 A_2 + i\kappa_{21} A_1(z), \tag{12.52b}$$

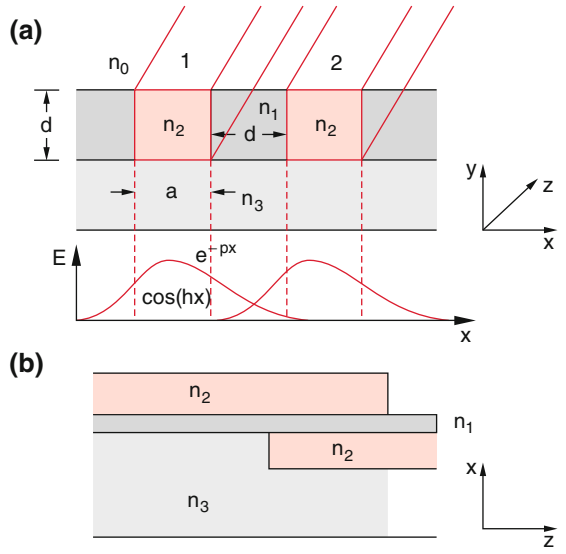


Fig. 12.42 Coupling between two close wave-guides by superimposing field distributions **a**) cross section in the *x*-*y*-plane **b**) in the *x*-*z*-plane

where $\beta = n_2 \cdot \cos\vartheta$ is the propagation parameter and κ is the coupling coefficient. If both waveguides have equal dimensions is $\beta_1 = \beta_2 = \beta$ and $\kappa_{12} = \kappa_{21} = \kappa$.

Wirth the initial conditions $A_1(0) = 1$ and $A_2(0) = 0$ the solutions of (12.52a, 12.52b) are

$$A_1(z) = \cos(\kappa z) \cdot e^{i\beta z}, \tag{12.53a}$$

$$A_2(z) = -i \sin(\kappa z) \cdot e^{i\beta z}, \tag{12.53b}$$

The power of the waves in the two waveguides is then (Fig. 12.42)

$$P_1(z) \propto A_1 \cdot A_1^* = \cos^2(\kappa z) \tag{12.54a}$$

$$P_2(z) \propto A_2 \cdot A_2^* = \sin^2(\kappa z) \tag{12.54b}$$

This shows that the energy oscillates back and forth between the two wave guides. After the path $z_1 = \pi/2\kappa$ it has been completely transferred from 1 to 2 and after $z_2 = \pi/\kappa$ it is again back to 1.

Choosing the coupling length just equal to z_1 the wave energy is completely transferred from waveguide 1–2; for $z = \pi/(4\kappa)$ half of the energy is transferred to 2.

Remark The problem is completely analog to that of two coupled pendula (see Vol. 1, Sect. 11.8).

If also other losses of the wave besides the coupling losses are taken into account (for instance absorption of the waveguide material) a complex refractive index is introduced (see the discussion in Sect. 8.1). The complex propagation parameter then becomes.

$$\beta = kn_2 \cdot \cos\vartheta = \beta_r - i\alpha/2$$

The wave propagating in the waveguide is a damped wave and one obtains instead of (12.54a and 12.54b) the equations

$$\begin{aligned} P_1(z) &= \cos^2(\kappa z) \cdot e^{-\alpha z} \\ P_2(z) &= \sin^2(\kappa z) \cdot e^{-\alpha z} \end{aligned} \tag{12.55}$$

12.7.4 Integrated Optical Elements

The large technical significance of integrated optics is owed to the combination of integrated light sources in thin film technology (semi-conductor lasers) integrated optical elements (lenses, prisms waveguides), micro-optical detectors and optical fibers [31, 32].

In this way source, communication channel and detector can be combined in integrated technique, i.e. small and compact. As example emitter, lens, waveguide and detector in effective planar layer structure are shown schematically in Fig. 12.43. The cylinder lens is realized by a region Δz with different refractive index, the semiconductor laser by a sandwich structure of silicon layers doped with different concentrations of impurity atoms (see Vol. 3, Sect. 8.4 and 14.3). The distribution of the input energy onto the different exit channels can then be realized by coupling adjacent wave guides, which can be controlled by electro-optical modulation (see foregoing section).

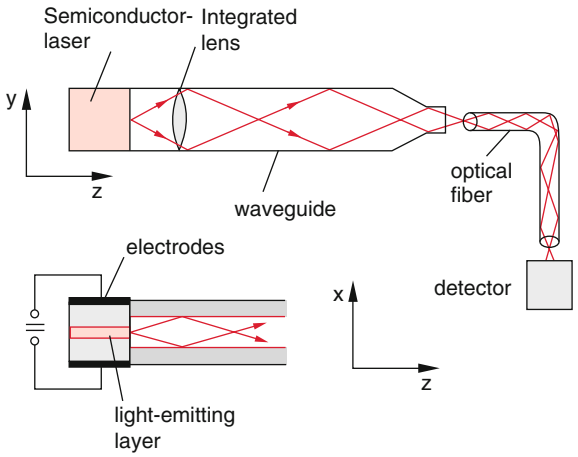


Fig. 12.43 Light source, wave guide optics and optical fiber with detector in integrated design

12.8 Optical Fibers

Optical fibers are thin fibers of fused silica with diameters between 5 and 50 µm, where a core zone has a slightly

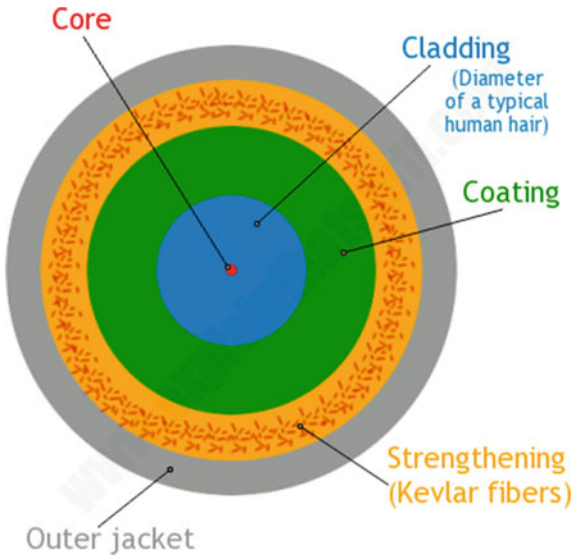


Fig. 12.44 Cross section of an optical fiber (www.explainthatstuff.com)

higher refractive index than the surrounding mantle called cladding (Fig. 12.44). The inner core has a refractive index which is larger than that of the surrounding cladding. The silicon cladding is covered by a plastic protective coating in order to protect the fiber from external damage. Typical values of the refractive indices are $n_{\text{core}} = n_1 = 1.48$; $n_{\text{cladding}} = n_2 = 1.46$. A light wave coupled into the core is trapped due to total reflection at the boundary between core and cladding as long as the angle between wave vector $\boldsymbol{k}$ and boundary is above the critical angle γ_g for total reflection with $\sin\gamma_g = n_2/n_1$ (Fig. 12.45) (see Sect. 8.4.6).

The maximum acceptance angle α_a for the coupling into the core is with $\sin\alpha/\sin\beta = n_1$ given by

$$\begin{aligned}\sin\alpha_{\max} &= n_1 \cdot \sin\beta_g = n_1 \cdot \cos\gamma_g = n_1 \cdot \sqrt{1-\sin^2\gamma_g} \\ &= n_1 \cdot \sqrt{1-(n_2/n_1)^2} = \sqrt{n_1^2 - n_2^2}.\end{aligned} \tag{12.56}$$

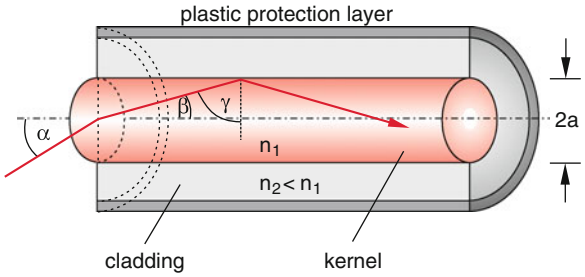


Fig. 12.45 Design of an optical fiber

The angular acceptance range

$$A_N = \sin\alpha_a = \sqrt{n_1^2 - n_2^2} \tag{12.57}$$

is the ***numerical Aperture*** of the optical fiber.

Example

$n_1 = 1.48$,
$n_2 = 1.46 \rightarrow \sin\alpha_{\max} = 0.14 \Rightarrow \alpha_{\max} = 8°$.

The numerical aperture limits the maximum light power that can be coupled into the fiber core. Using a lens with focal length f to focus a light beam with diameter d onto the core (Fig. 12.46) we get

$$\tan\alpha_{\max} = d/2f$$

The advantage of optical fibers is the possibility of bending the fiber, for instance winding a long piece of the fiber on a circle with many windings. As long as the radius of curvature of a bent fiber is not too small, the light is still trapped in the core (Fig. 12.47).

The radial profile of the refractive index n in optical fibers can be chosen in different ways: For the *step index fiber* (Fig. 12.48a) is $n(r) = \text{const}$ within the core. At the

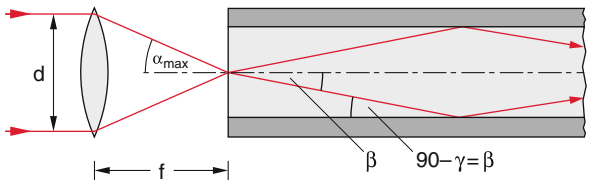


Fig. 12.46 Coupling of light into an optical fiber

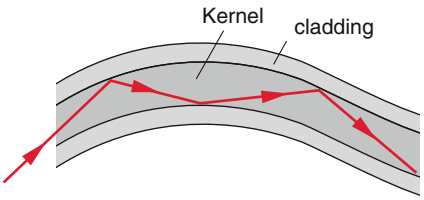


Fig. 12.47 Propagation of light in an optical fiber with step-index profile with total reflection at the interface between core and cladding

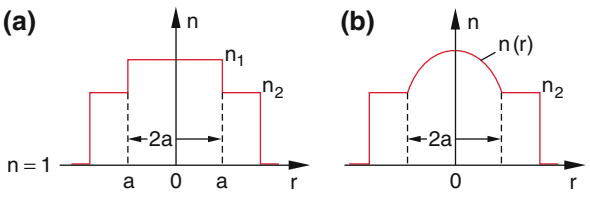


Fig. 12.48 Radial index profile **a**) for a step-index profile fiber **b**) for a gradient index fiber

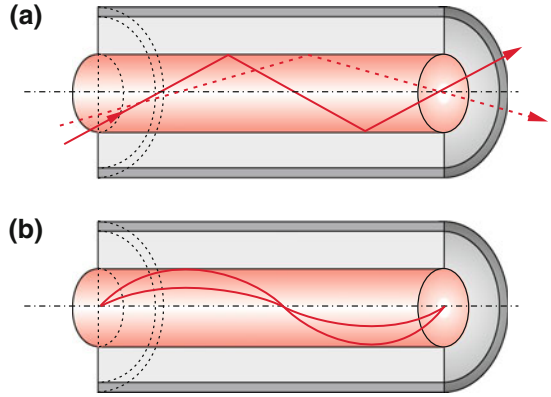


Fig. 12.49 Propagation of different modes **a**) in a step-index fiber **b**) in a gradient index fiber

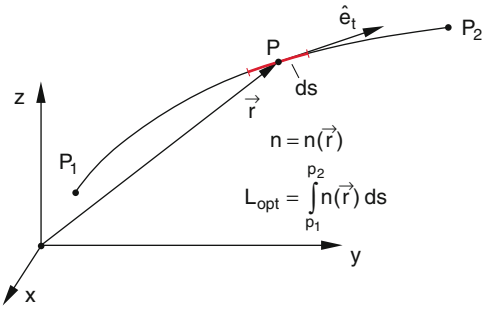


Fig. 12.50 Application of Fermat's principle to light propagation in a medium with locally varying refractive index $n(\boldsymbol{r})$

boundary core-mantle n makes a step from n_1 to n_2 and is then again constant in the mantle. In the gradient index fiber $n(r)$ has a radial profile as shown in Fig. (12.48b). Often $n(r)$ has a parabolic profile $n(r) = n(0) - b \cdot r^2$. The light rays in the core of gradient index fibers are curved (Fig. 12.49b). The advantage of gradient index fibers is the smaller dependence of the propagation velocity from the entrance angle (see next section).

12.8.1 Light Propagation in Optical Fibers

In Sect. 7.9 we have discussed the propagation of different wave modes in waveguides with rectangular cross section. In a similar way also in waveguides with circular cross section the propagation characteristics depends on the specific mode of the wave and on the dimensions of the waveguide. For instance, in a step index fiber with core diameter $2a < \lambda$ only the fundamental mode TEM_{00} can propagate.

The propagation velocity of a wave in an optical fiber differs for the different transverse modes. The difference is smaller for step index fibers than for gradient index fibers (Fig. 12.49). Therefore generally either step-index fibers or even better single mode fibers are used for optical communication. We will discuss the propagation characteristic in such fibers in more detail [32] (Fig. 12.50).

According to Fermat's principle (see Sect. 9.1) the light chooses its path between two points P_1 and P_2 such that the optical path length

$$L_{\text{opt}} = \int_{P_1}^{P_2} n(r)\,\mathrm{d}s = \text{Minimum} \tag{12.58}$$

becomes minimum (Fig. 12.50). With $\mathrm{d}s = \hat{e}_t \cdot \mathrm{d}\boldsymbol{r} \Rightarrow \mathrm{d}L_{\text{opt}} = n(\mathbf{r}) \cdot \hat{e}_t \cdot \mathrm{d}\boldsymbol{r}$ where $\hat{e}_t$ is the tangent unit vector at the point P. On the other hand is $\mathrm{d}L_{\text{opt}} = (\mathbf{grad} L_{\text{opt}}) \cdot \mathrm{d}\boldsymbol{r} \Rightarrow \nabla L_{\text{opt}} = n(\boldsymbol{r}) \cdot \hat{e}_t$. Scalar multiplication of $\mathbf{ds}$ with $\hat{e}_t$ gives $\hat{\boldsymbol{e}}_t \mathrm{d}s = \mathrm{d}\boldsymbol{r}$ and therefore

$$\nabla L_{\text{opt}} = n(\boldsymbol{r}) \cdot \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}s} = n(\boldsymbol{r}) \cdot \hat{e}_t \tag{12.59}$$

Differentiation of (12.59) with respect to $\mathrm{d}s$ yields because $\mathrm{d}\hat{\boldsymbol{e}}_t/\mathrm{d}s = 0$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(n \cdot \frac{\mathrm{d}\boldsymbol{r}}{\mathrm{d}s}\right) = \frac{\mathrm{d}}{\mathrm{d}s}\left(\nabla L_{\text{opt}}\right) = \nabla n(\boldsymbol{r}) \tag{12.60}$$

For gradient index fibers often a parabolic index profile is chosen

$$\begin{aligned} n(r \le a) &= n_1\left(1 - \Delta\left(\frac{r}{a}\right)^2\right) \\ n(r \ge a) &= n_2 \\ \text{mit } \Delta &= (n_1 - n_2)/n_1. \end{aligned} \tag{12.61}$$

where $\Delta = (n_1 - n_2)/n_1$ is the relative difference of the refractive indices n_1 of the core and n_2 of the cladding.

Since n depends solely on the distance r from the fiber axis, is $\nabla n = \mathrm{d}n/\mathrm{d}r$ and we obtain from (12.60) the equation in cylinder coordinates (r, φ, z) (see problem 12.11)

$$\frac{\mathrm{d}^2 r}{\mathrm{d}z^2} = \frac{1}{n(r)}\frac{\mathrm{d}n}{\mathrm{d}r} \tag{12.62}$$

Inserting (12.61) finally yields the equation

$$\frac{\mathrm{d}^2 r}{\mathrm{d}z^2} + \frac{2\Delta}{a^2} r = 0 \tag{12.63}$$

with the solution

$$r(z) = a \sin\left(\frac{\sqrt{2\Delta}}{a} \cdot z\right). \tag{12.64}$$

The path $r(z)$ of the light rays in a gradient index fiber shows a sinusoidal pathway around the axis $r = 0$ (Fig. 12.49b) with the period

$$\Delta z = \Lambda = 2\pi a/\sqrt{2\Delta}. \tag{12.65a}$$

Introducing the wavenumber

$$K = 2\pi/\Lambda = \frac{1}{a}\sqrt{2\frac{n_1 - n_2}{n_1}}. \tag{12.65b}$$

The velocity of the light propagating along the trajectory (12.64) is

$$\begin{aligned} v_{\mathrm{ph}}(z) &= \frac{c}{n(z)} = \frac{c}{n_1\left(1 - \Delta \sin^2 Kz\right)} \\ &= \frac{c}{n_1\left(1 - \frac{K^2}{2}a^2 \sin^2(Kz)\right)} \end{aligned} \tag{12.66}$$

The phase velocity v_{ph} oscillates between the minimum value c/n_1 for $K \cdot z = n \cdot \pi$ (which occurs for $r = 0$) and the maximum value c/n_2 for $K \cdot z = (n + \frac{1}{2})\pi$ which is reached for $r = a$ i.e. at the boundary between core and cladding.

The propagation velocity in the z-direction depends on the incidence angle α. Its mean value averaged over one period Λ is

$$\langle v_z \rangle = \frac{\Lambda}{T} = \frac{\Lambda}{\int_0^{\Lambda} \frac{\mathrm{d}z}{v_{\mathrm{ph}} \cdot \cos\alpha}} \tag{12.67}$$

$$\langle v_z \rangle = \frac{c}{n_1}\left(1 - \left(\frac{\Delta \cdot \alpha}{2\alpha_0}\right)^2\right), \tag{12.68}$$

where $\alpha_0 = \sqrt{(2\Delta)} \cdot (\mathrm{d}r/\mathrm{d}z)_{r=0}$ is the angle between the trajectory and the symmetry axis $r = 0$.

Since $\Delta \ll 1$, Eq. (12.68) shows that the travel time of light through gradient index fibers depends less strongly on the incidence angle α than in step index fibers (Fig. 12.49a).

12.8.2 Absorption in Optical Fibers

The losses which light suffers when propagating through optical fibers are of fundamental importance for optical communication over large distances. They are caused by absorption, scattering and leakage from the core into the cladding. Their combined effect is called the *fiber attenuation*.

If the relative power loss on the fiber length $\mathrm{d}L$ is

$$\frac{\mathrm{d}P}{P} = -\kappa \cdot \mathrm{d}L, \tag{12.69}$$

Integration yields

$$P(L) = P(0) \cdot \mathrm{e}^{-\kappa L}. \tag{12.70}$$

The transmitted power decreases exponentially with increasing propagation length L. The damping constant

$$\kappa = -\frac{1}{L}\ln\frac{P(L)}{P(0)} \tag{12.71}$$

depends on the fiber material and the wavelength λ. Since the damping constant can vary over several decades, in telecommunication generally the decadic logarithm is used and the damping coefficient is defined as

$$\alpha = -\frac{10}{L}\log\frac{P(L)}{P(0)} \tag{12.72a}$$

with the unit decibel (db) per km fiber length. Rearranging gives

$$P(L) = P(0) \times 10^{-\alpha L/10} \tag{12.72b}$$

Example

For $\alpha = 0.5\,\mathrm{dB/km}$ the transmitted power has decreased after a fiber length of 10 km to $P(L)/P(0) = 10^{-0.5} = 0.316 \cdot P(0)$. The power has decreased to 31.6% of its initial value.

For $\alpha = 0.1$ db/km (which can be realized today) the transmitted power decreases after 10 km to 80% and after 100 km to 10% of its initial value.

In Fig. 12.51 the spectral damping curve of a modern optical silicon fiber is shown. It shows that the damping coefficient has a minimum at $\lambda = 1.6\mu\mathrm{m}$. This is due to the superposition of several effects: The scattering cross section for Rayleigh scattering is proportional to $1/\lambda^4$ and increases therefore steeply with decreasing wavelength (see Sect. 8.2). The absorption is essentially caused by the low-wavelength side of the infrared absorption by excitation of vibrations which impurity atoms perform against the lattice of the fiber

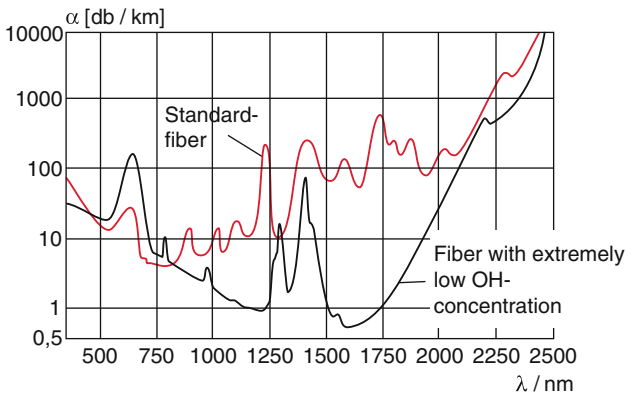


Fig. 12.51 Attenuation $\alpha(\lambda)$ of a standard fiber, which contains small residual impurities (red curve) and for a fiber with very low OH-concentration

material. Scattering and absorption increase drastically with the concentration of impurity atoms and molecules. In particular the OH-radical absorbs strongly around $\lambda = 1.4\,\mu m$. Therefore it is essential to use extremely purified material. For shorter wavelengths the long wavelength side of the UV-absorption due to electronic transitions of the atoms and molecules becomes important [32].

12.8.3 Optical Pulse Propagation in Fibers

The propagation of optical pulses in fibers is not only influenced by absorption and scattering but also by dispersion which alters the form of the pulses. Dispersion has two different causes:

(a) The different propagation modes of waves in the fiber (Fig. 12.49) have different velocities (mode dispersion, see Eq. 7.47)
(b) The refractive index n and the phase velocity $c' = c/n(\lambda)$ depends on the wavelength λ (Fig. 12.55). Since optical pulses with the temporal width ΔT are composed of an infinite number of waves with frequencies ν within the frequency range $\Delta\nu = 1/\Delta T$, the different velocities of these frequencies result in a change of the pulse form $I(t)$ particular in a broadening of the pulse. Since the pulse width has to be smaller than the time interval between successive pulses this effect limits the fiber length for the optical communication for a given pulse rate (Fig. 12.52).

In an optical fiber with a core diameter $2a \gg \lambda$ generally many propagation modes are possible. This has been illustrated in Sect. 7.9 by the example of a waveguide and in Fig. 12.49 for two different modes in an optical fiber. Therefore multimode fibers are not appropriate for optical communication over large distances. On has to use monomode fibers with core diameters $2a < 3\lambda$. This demands a very precise adjustment for coupling the light into the fiber. This is, however, nowadays technical feasible, even for the splicing of fiber ends under operating conditions in outside cable trenches.

The dispersion $n(\lambda)$ is shown in Fig. 12.53 for a silicon fiber doped with GeO_2 which is often used for optical communication.

The propagation of pulses is governed by the group velocity (see Sect. 8.2 and Vol. 1, Sect. 11.x)

$$v_G = v_{ph} + k\frac{dv_{ph}}{dk} = \frac{c}{n_r + \omega\, dn_r/d\omega} \tag{12.73}$$

which depends on the refractive index $n = n_r - i\kappa$ (Sect. 8.1). For the pulse deformation the dispersion of the group velocity $dv_G/d\lambda$ resp. dv_G/dk is responsible. Figure 12.53 shows that for standard optical fibers the group velocity has a maximum for $\lambda \approx 1.3\,\mu m$ and therefore its dispersion becomes zero. This wavelength is therefore optimal for the transmission of high bit rates. The absorption, however, has a minimum at $\lambda = 1.5\,\mu m$. This implies that the minimum dispersion is paid dearly for a higher absorption. It is possible to shift the maximum of v_G by increasing the concentration of impurity atoms to $\lambda = 1.5\,\mu m$, but this leads to a slightly higher absorption than for a lower concentration (Fig. 12.54).

The question is now whether it might be possible to find a better solution. This search has been indeed successful by making use of nonlinear effects which allow the generation of pulses called **solitons**, that do not change their form during their propagation through the fiber.

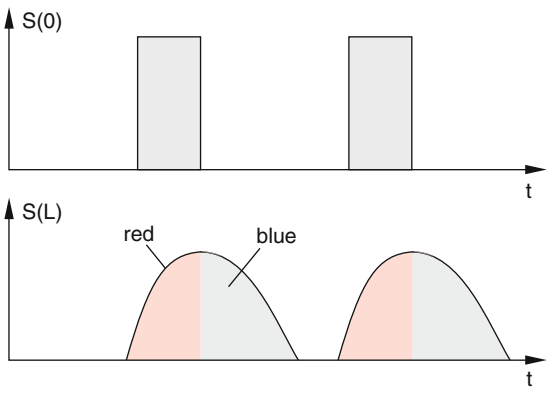


Fig. 12.52 Broadening and frequency shift (chirp) of an optical pulse propagating through a medium with normal dispersion

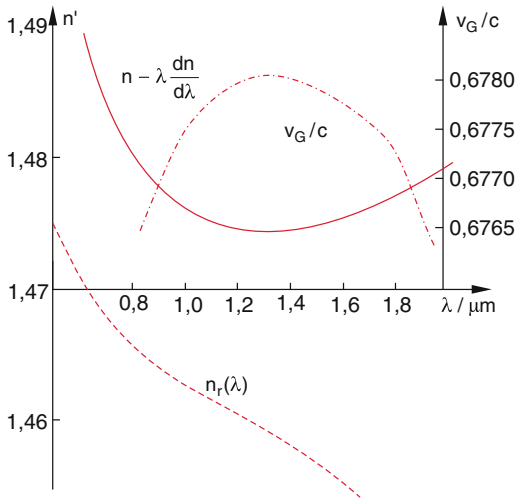


Fig. 12.53 Refractive index $n(\lambda)$, group refractive index $n - \lambda \cdot dn/d\lambda$, and group velocity v_G of a quartz fiber doped with 7 mole% GeO_2

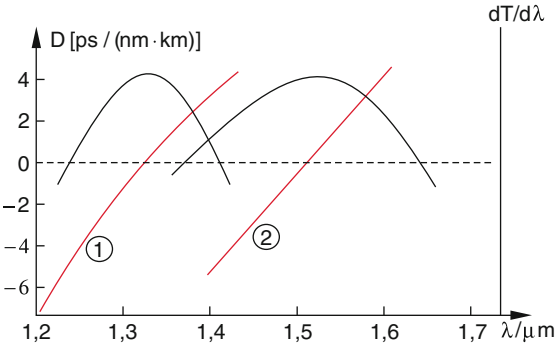


Fig. 12.54 Group velocity dispersion (Black curves, left ordinate) and transit time dispersion (red curves 1 and 2, right ordinate) for two different optical fibers with different germanium doping

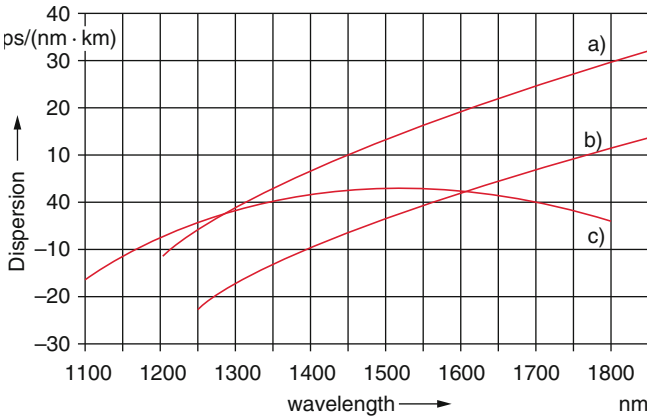


Fig. 12.55 Dispersion of some fiber-types. **a**) Standard single-mode fiber **b**) dispersion-shifted single mode fiber **c**) dispersion-smoothed single mode fiber

12.8.4 Nonlinear Pulse Propagation; Solitons

For a sufficiently high intensity of the light pulses the atomic electrons are induced to oscillations with amplitudes that exceed the harmonic range (linear restoring force). The refractive index is then no longer described by Eq. (8.1a) but depends on the intensity. We can write:

$$n(\omega, I) = n_r(\omega) + n_2 \cdot I, \qquad (12.74)$$

where the second term $n_2 \cdot I$ becomes comparable to the first term only for high intensities. The phase $\phi = \omega t - kz$ of an optical wave $E = E_0 \cos(\omega t - kz)$ is with $k = 2\pi/\lambda = n \cdot \omega/c$

$$\phi = \omega t - \omega \cdot n \cdot z/c = \omega(t - n_r z/c) - A \cdot I(t), \quad (12.75a)$$

where $A = n_2 \omega z/c$. It therefore depends on the intensity

$$I(t) = c \cdot \varepsilon_0 \int |E_0(\omega, t)|^2 \cos^2(\omega t - kz) \mathrm{d}\omega \qquad (12.75b)$$

The momentary optical frequency of the wave which is equal to the time derivative of the phase

$$\omega = \mathrm{d}\phi/\mathrm{d}t = \omega_0 - A \cdot \frac{\mathrm{d}I}{\mathrm{d}t} \qquad (12.75c)$$

depends on the temporal change of the intensity.

When a light pulse with center frequency ω_0 and intensity $I(t)$ propagates through the fiber, the intensity derivative at the leading edge of the pulse is $\mathrm{d}I/\mathrm{d}t > 0$ and therefore $\omega < \omega_0$ while at the trailing edge $\mathrm{d}I/\mathrm{d}t < 0$ and therefore $\omega > \omega_0$. This frequency variation during the pulse duration is called **frequency chirp**. It leads to a spectral broadening of the pulse and also to a temporal broadening, because the red-shifted frequencies at the leading edge precede the blue-shifted frequencies. If one now chooses the optical wavelength λ within the spectral range of "anomalous dispersion" with $\mathrm{d}n/\mathrm{d}\lambda > 0$ the red frequencies have a lower phase velocity then the blue ones. For the correctly chosen intensity the two opposite effects just compensate. This gives pulses which are spectrally broadened but have a temporal pulse width that stays constant. Such pulses are called **solitons** [40].

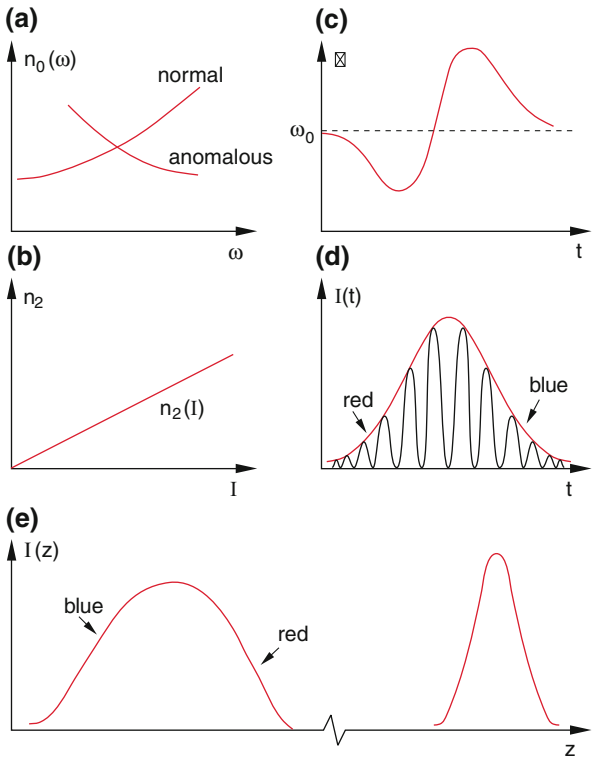


Fig. 12.56 **a**) Normal and anomalous dispersion **b**) $n_2(I)$ **c**) frequency chirp caused by the nonlinear contribution $n_2 \cdot I(t)$ **d**) temporal progression of a light pulse with frequency chirp **e**) spatial form of a pulse, which becomes shortened after a path length $z(I)$ in a medium where the normal and anomalous dispersion have been compensated and which travels with a constant pulse profile further on (soliton)

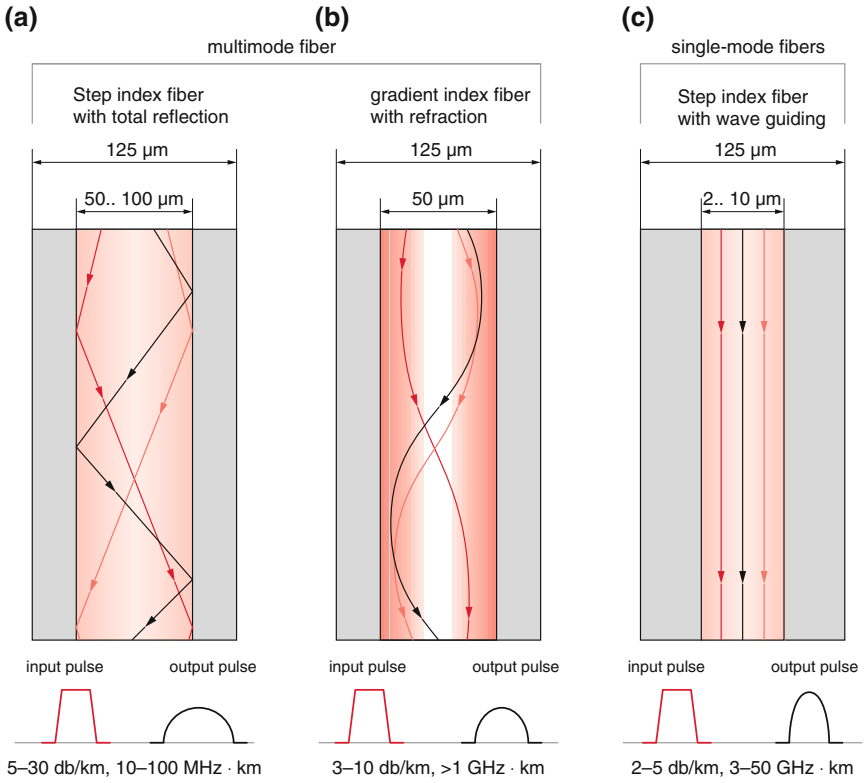


Fig. 12.57 Light propagation **a**) in a step index fiber **b**) in a gradient index fiber **c**) in a single mode fiber. Below the figures are shown incident and exit pulse profile, the attenuation in db/km, and the achievable bandwidth in MHz km

This is again illustrated in Fig. 12.56. The linear dispersion causes a temporal broadening of the pulses, while the nonlinear dispersion increases their spectral width. In the spectral range of anomalous dispersion the temporal pulse width of the broadened pulse decreases again. In Fig. 12.57 the change of the pulse form and width are illustrated for the different fiber types.

This example shows that for the technical realization of new methods a sound fundamental research is required in order to achieve optimum solutions [33, 34] (Fig. 12.55).

12.9 Optical Communication

Optical signal transmission over larger distances has been used for many centuries for fast communication of important facts in particular for military communication. One possibility is the system developed 1792 by Claude Chappe (Fig. 12.58), where the letters of the alphabet were encoded as special positions of the arms of the transmitting station. These positions could be viewed by the next station about 10–20 km away. A whole network of stations was built in France. Since the mechanical change of the arm positions took some time, the communication was not very fast.

About 100 years later the first electrical communication technique, where voltage- or current signals were sent through electrical cables were invented. It remained for many years the only way to communicate information with high speed over large distances. Later on the wireless communication and radio transmission were developed (Fig. 12.59).

Recently the digital optical communication has gained increasing importance, where short light pulses are send through optical fibers. The large bandwidth of this technique allows the simultaneous transmission of several thousand channels. With digital communication systems meanwhile bitrates of up to 10^{12} bits/s with optical pulse widths below 10^{-12} s have been reached. Another advantage of this method is its security against interception. The attenuation of the optical pulses by fiber losses is smaller by several orders of magnitude than for electric signals sent through copper cables (see Sect. 12.8.3 and 12.8.4). Since optical communication is a very important part of modern optics we will discuss at the end of this textbook the required system

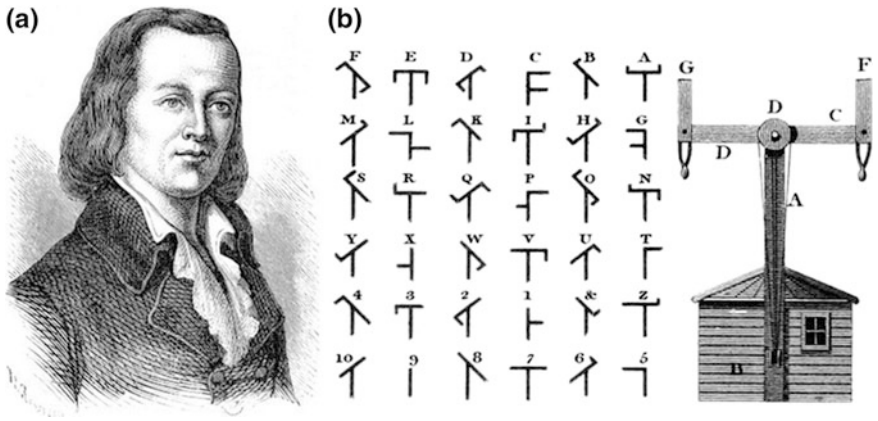


Fig. 12.58 Claude Chappe and his Morese code which can be mechanically realized by moving the arms of the transmitting station

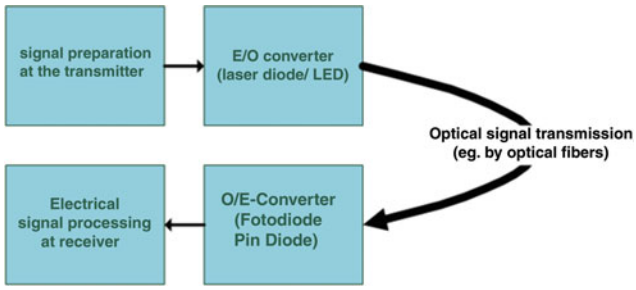


Fig. 12.59 Signal processing and transmission for optical communication

components and the merits and drawbacks of this exciting communication technique and also some problems which have still to be solved (Fig. 12.59).

The principle of information transmission is schematically depicted in Fig. 12.60. The input information is processed in the transmitter and is sent through the transmission line. The receiver at the end of this line processes and amplifies the signals, filters the wanted information and converts it into a form readable by the recipient.

The system for optical communication is depicted in Fig. 12.61 in more detail. The output beam of a light source (generally a laser) is amplitude- or phase modulated by the signal which should be transmitted and is then sent through the transmitter optics (generally an optical fiber). Since many

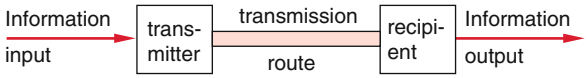


Fig. 12.60 Basic principle of information transfer

different channels can be simultaneously transmitted, a channel selector separates the different channels and the detector following the receiver optics converts the optical signal into an electrical output which gives the output information. The semiconductor laser as the source emits very short pulses and the repetition rate of the pulses is modulated by the signal. A fast photodiode receives the pulses and converts them into electrical pulses. The information is presented in digital form as binary code and is extracted by the receiver and transformed into an analogue signal for example as music or language text.

In Fig. 12.62 the general scheme for optical communication is shown as a flowchart.

The great advantage of optical communication through fibers is the available large bandwidth. For a wavelength $\lambda = 1.5\,\mu m$ (this corresponds to the optical frequency $\nu = c/\lambda = 2 \times 10^{14}\,s^{-1}$) about 2 million channels width a bandwidth of 100 MHz each can be simultaneously transmitted through a single optical fiber. De facto however, this number is considerably lower because the dispersion of the fiber broadens the pulses and limits the maximum pulse rate.

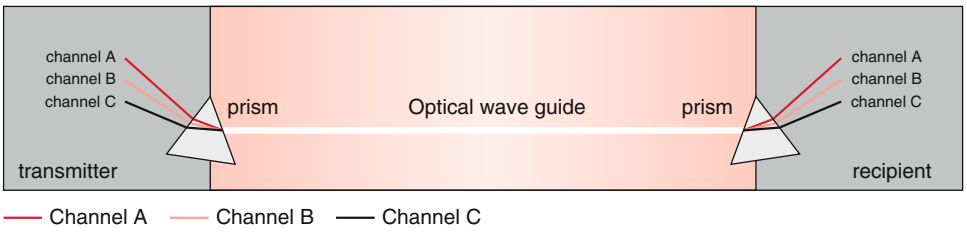


Fig. 12.61 Simultaneous transfer of several channels achieved by optical wavelength multiplexing (http://Einstein.informatik.uni-oldenburg.de/rechnernetze/optische.htm)

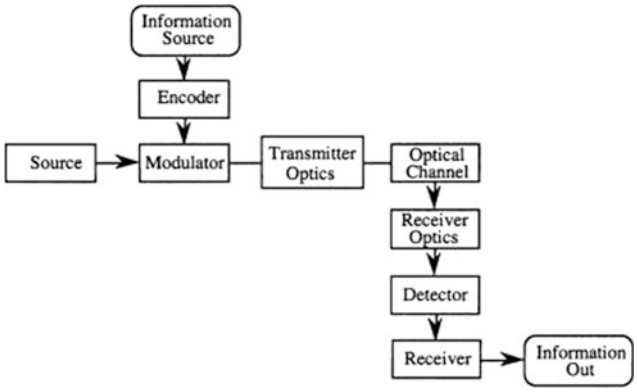


Fig. 12.62 More detailed block diagram of optical information communication

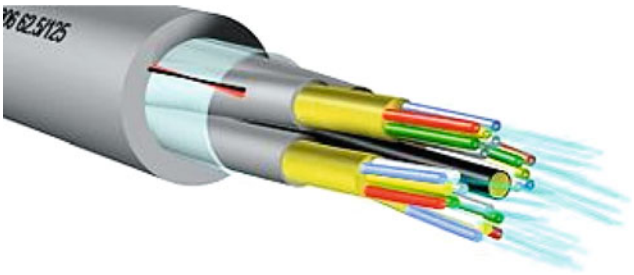


Fig. 12.63 Optical glass fiber bundle with protective jacket

Example

With a pulse width $\tau = 7.5\,\text{ps}$ the Fourier-limited spectral width of the pulse is $\Delta\nu \approx \frac{1}{\tau} = 1.4 \times 10^{11}\,\text{s}^{-1} = 140\,\text{GHz} \Rightarrow \Delta\lambda = (c/\nu^2)\Delta\nu = 2\,\text{nm}$ for $\nu = 2 \times 10^{14}\,\text{s}^{-1}$. With this spectral width the dispersion is about 2 ps/km. After 50 km fiber length the pulses have broadened to 100 ps. The maximum pulse rate is then limited to 10^{10} pulses/s and the maximum transmitted bandwidth is 10 GHz. The number of simultaneously transmitted channels is now $N = 10^{10}/10^8 = 100$.

Most connections between two large cities use fiber bundles which contain many single fibers (Fig. 12.63). This multiplies the number of channels that can be simultaneously transmitted.

More detailed information can be found in the excellent book by Rogers [36] and in [35–40]

Summary

- Confocal microscopy allows a high spatial resolution in the x-y-focal plane perpendicular to the light propagation as well as in the z-direction. The achieved contrast is generally higher than in conventional microscopy.
- With the near field microscopy a spatial resolution below 100 nm can be achieved. One needs, however, intense light sources, such as lasers for the illumination of the object. This method is mainly used for inspection of structures on surfaces.
- Active optics corrects unwanted deformations of mirrors in astronomical telescopes by electronically controlled elements. It minimizes image aberrations, improves the quality of the image and increases the angular resolution of the telescope.
- Adaptive optics corrects the degradation of the image, caused by turbulences in the atmosphere. In combination with active optics it achieves a nearly ideal image quality of astronomical mirrors where the image of a star is equal to the diffraction disc of a point source limited by the diameter D of the primary mirror of the telescope. The angular resolution approaches $\Delta\alpha = \lambda/D$.
- Diffraction optics utilizes diffraction and interference for the imaging of objects (Fresnel lens) or for deflection of light beams (step plate). This technique allows the fabrication of micro-lenses and lens arrays in integrated design.
- Holography allows the construction of a three-dimensional image of an object. It is based on the interference of the signal wave scattered by the object with a coherent reference wave. This enables the measurement of the phases of signal waves scattered by different points of the object. The information about the object is stored in encoded form in the hologram. The illumination of the hologram with a reconstruction wave produces the real three-dimensional image of the object.
- The combination of holography and interferometry allows the visualization of small temporal changes of an object or of the deviation of objects from a reference model.
- Three-dimensional holograms stored in special materials can be used as storage for information with high packing density.
- Fourier-optics is based on the insight that for Fraunhofer diffraction the amplitude distribution in the diffraction plane is equal to the Fourier-transform of the light amplitude in the object plane. The further imaging of the diffraction plane into the image plane gives the real image of the object.
- Manipulations of the diffraction pattern (optical filtering by apertures, optical amplitude filters, phase plates or holograms) can specifically alter the image of an object. If only low spatial frequencies are transmitted in the diffraction plane finer details in the image disappear (low pass filter). If only high spatial frequencies are transmitted finer details of the object are imaged with higher contrast (high pass filter).
- By high pass filtering phase objects (smear formation or turbulences in liquids) can be made visible.
- Integrated optics uses microscopic small optical waveguides for modulation and deflection of light waves. They are produced by integrated techniques (etching, evaporation techniques or the use of microscopic masks). This allows the integration of light source, wave guide and detector on a single chip.
- Fast optical communication is based on the transmission of ultrashort laser pulses through optical fibers with very small damping. This allows the realization of long transmission distances. The maximum possible bit rate is limited by the dispersion of the fiber. The dependence of the refractive index on the intensity of the transmitted light makes it possible that under optimum conditions optical pulses can be transmitted without changing their form $I(t)$ during the passage through the fiber (solitons).

Problems

12.1 For a special application of confocal microscopy the diameter of the circular aperture is 0.01 mm, the distance aperture-lens is 100 mm, the focal length of the lens f = 10 mm

(a) Where is the focus located and what is its diameter derived for geometrical optics and if diffraction is taken into account?

(b) What is the distance Δz from the focal plane, where the intensity of the light scattered by the sample and transmitted through the aperture has decreased to ½ of its maximum value for $\Delta z = 0$?

12.2 A star as point source is imaged by a parabolic mirror with diameter D = 5 m

(a) What is the diameter of the central diffraction disc?

(b) The mirror is deformed in such a way that its surface can be described by $y^2 = 4f\,\varepsilon x$ with $|\varepsilon - 1| \ll 1$ instead by (9.11). How large is now the image of the star?

12.3 A star has the zenith distance $\zeta = 60°$.The refractive index n averaged over the observation time has the mean fluctuation $\delta n = 3 \times 10^{-2} n$ with n = 1.00027. How large is the angular broadening and what is the size of its image for a focal length f = 10 m of the mirror?

12.4 Rectangular parallel grooves (depth $h = 1\,\mu m$, width $b = 2\,\mu m$, distance between the grooves $d = 4\,\mu m$) are etched into a glass plate (n = 1.4). Under which angles can transmitted light with the wavelength $\lambda = 500\,nm$ be observed, when it falls onto the glass plate as parallel beam

(a) vertical

(b) with the incidence angle $\alpha = 30°$ against the surface normal?

12.5 A Fresnel lens with focal length f = 10 mm and diameter D = 20 mm should be realized

(a) How large must the radii of the circular grooves be (depth 1 µm)? How many grooves are possible?

(b) Could a lens with D/f = 2 also be realized as refractive lens?

(c) How could the Fresnel lens be technically produced?

12.6 A holographic grating with 10^5 parallel grooves and a groove distance $d = 1\,\mu m$ shall be produced by illumination of a photo layer by two plane waves. How large should be the diameter of the enlarged beams and what is the angle between the two wave vectors for symmetric illumination?

12.7 What is the amplitude distribution of 5 light sources in the diffraction plane of a lens with focal length f for the positions $(x_0, 0)$, $(-x_0, 0)$, $(0, -y_0)$, $(0, +y_0)$ and $(0, 0)$ of the light sources

(a) for all 5 light sources

(b) if the sources $(x_0, 0)$ and $(-x_0, 0)$ are extinguished

(c) if the source $(0, 0)$ is extinguished

(d) if all sources except $(0, 0)$ are extinguished?

12.8 A parallel light beam is incident onto a grating with parallel grooves and bars (width $b = 1\,\mu m$, distance $d = 2\,\mu m$)

(a) What is the far field amplitude- and intensity distribution in the diffraction plane?

(b) How is the distribution altered if only every 3rd groove is open?

12.9 A planar wave guide in z = direction has the width $a = 2\,\mu m$ and the refractive index n_2 = 2. What are the three lowest modes with the mode numbers m_s = 1, 2, and 3 for $\lambda = 500\,nm$? What is the minimum difference $\Delta n = n_2 - n_1$ for keeping these modes within the waveguide? Which angles ϑ have the k-vectors of these modes against the z-direction? What are their parameters p, h, q?

12.10 A light pulse (λ = 1.3 µm) with the width $\Delta t = 1\,ps$ propagates through a single mode fiber with refractive index n = 1.5 and the dispersion $dn/d\lambda = 2 \times 10^{-6}/nm$. After which propagation length has the pulse width doubled due to dispersion?

12.11 Derive (12.62) from (12.60) for the case of the parabolic index profile (12.61).

12.12 Calculate the dependence of the mean propagation velocity on the incidence angle α in a step index fiber and a gradient fiber. Show, that the maximum value of α in a gradient fiber is given by $\alpha_{max} = \sqrt{2}\Delta$ with Δ defined in Eq. (12.61)

References

1. W. Gray Jerome, R.L. Price: Basic Confocal Microscopy, 2nd ed. (Springer, Heidelberg 2018)
2. St. W Paddock: Confocal Microscopy (Springer, Heidelberg 2014)
3. C. Shephard P.M. Shotton: Confocal Laser scanning Microscopy Microscopy Handbook Series Vol 38, (Springer, Heidelberg, 1997)
4. J. Toporski, Th, Deing, O. Hollricher: Confocal Raman Spectroscopy (Springer Heidelberg 2018)
5. M. Paesler, P.J. Moyer : Near Field Optics (John Wiley & Sons New York 1996)

6. D. Courion; Near Field Microscopy and Near Field Optics (Imperial College Press London 2003)
7. J.P Fillard: Near Fiedl Optics (World Scientifique Singapore 1996)
8. L. Novotny, B. Hecht: Principles of Nano-Optics (Cambridge University Press Cambridge 2006)
9. Chanan, G. A.; Nelson, J. E.; Ohara, C. M.; Sirko, E. "Design Issues for the Active Control System of the California Extremely Large Telescope (CELT)". Proc. SPIE 4004, 363 (2000)
10. Chanan, G.; Troy, M.; Dekens, F.; Michaels, S.; Nelson, J.; Mast, T.; Kirkman, D. "Phasing the mirror segments of the Keck Telescopes: the broadband phasing algorithm," Applied Optics 101, 140 (January 1998)
11. A. Glindemann: Priniciples of Stellar Interferometry (Springer, Heidelberg 2011)
12. R.K. Thyson: Principles of Adaptive Optics (CRC Press, 4th ed. D. M. Alloin: Adaptive Optics for Astronomy (Nato Science Series C, Vol. 423, Springer Netherland 1994)
13. Joseph M. Geary: Introduction to Optical Testing (Photo-Optical 1993)
14. G. Saxby: Practical Holography (Taylor and Francis 4th Ed. 2017)
15. W. Davis: Holography: Techniques and Applications
16. P. Hariharan: Basic Hoolography Cambridge Universitý Press 2002
17. F. Unterseher, B. Schlesinger: Holographic Handbook (Ross Books 2010)
18. https://en.wikipedia.org/wiki/Holography
19. Pramond Rastogi: Holograsphic Interferometry (Springer Series in Optical Science Vol. 68 Heidelberg 2013)
20. S. Hirsch, P. Hering et.al.: Single-pulsed digital holographic topometry. Proc. SPIE 663101 (2007)
21. H.M. Smith: Holographic Recording Materials (Springer Topics in Applied Physics Vol. 20 1977)
22. E.G. Steward: Fourier Optics. An Introduction. (Dover Publ. 2nd ed. 2004)
23. J. Goodman: Introduction to Fourier Optics (W.H. Freeman 2017)
24. Francis T.S., Yu. S. Jutamulia: Optical Pattern Recognition (Cambridge Univ. Press 1998)
25. S. Theodoridis, K Koutroumbas: Pattern Recognition. (Elsevier, Oxford 2008)
26. W. Karthe, G. Wegner, H.D. Bauer, H.O. Moser: Integrated Optics and Micro-optics with Polymers (Vieweg-Teubner 1993)
27. H. Zappe, Claudia Duppé: Tunable Micro-Optics (2016) (e-book, free download)
28. R Winston, J.C. Minano P.G. Benitez: *Nonimaging Optics* (Academic, 2005)
29. N.F. Borrelli, *Microoptics technology: fabrication and applications of lens arrays and devices*. Marcel Dekker, New York (1999)
30. K. Okamoto: Fundamentals of Optical Waveguides (Academic Press, 2nd ed. 2010)
31. M.J. Adams: Introduction to Optical Waveguides (John Wiley & Sons, e-book)
33. J.A. Buck: Fundamentals of Optical Fibers (Wiley Interscience 2004)
    Jeff Hecht: Understanding Optical Fibers (Laser Light Press 2015)
    Vivek Altwayn: Fiber Optic Technology (Cisco Press 2004)
34. Akira Hasegawa: Optical Solitons in Fibers (Springer, Heidelberg 1990)
35. Reinhold Noé: Essentials of Modern Fiber Communication (Springer, Berlin 2016)
36. John, Gowar: Optical Communication Systems (Prentice Hall)
37. John Senior: Optical Fiber Communication: Principles and Practice Pearson 2008)
38. R. K. Singh: Fiber Optical Communication Systems. (Wiley 2012)
39. Gowind Prasant Agrawal: Fiber Optical Communication System (Wiley Series in Microwave and Optical Engineering, Vol. 1 (2010)
40. P.G.Drazin, R.S.Johnson: Solitons, An Introduction Cambridge Univ. Press 1989
    M. Remoissenet: Waves Called Solitons Springer, Heidelberg (2010)

Solutions of Problems

Chapter 1

1.1 The number of Na-atoms in each ball is

$$N = \frac{M}{m} = \frac{10^{-3}}{2.9 \cdot 1.67 \times 10^{-27}} = 2.06 \times 10^{22}$$

$\Longrightarrow$ The charge is

$$\begin{aligned} Q &= +e \cdot 0.1 \cdot 2.6 \times 10^{22} \\ &= 2.6 \cdot 10^{21} \cdot 1.6 \times 10^{-19}\,\mathrm{C} \\ &= 4.16 \times 10^{2}\,\mathrm{C}. \end{aligned}$$

The volume of each ball is

$$V = \frac{m}{\varrho} = 1.03\,\mathrm{cm}^3 = \frac{4}{3}\pi r^3$$

$$\Rightarrow r = \left(\frac{3 \cdot 1.03}{4\pi}\right)^{1/3} \mathrm{cm} = 0.63\,\mathrm{cm},$$

The surface is

$$S = 4\pi r^2 = 4.93\,\mathrm{cm}^2,$$

The surface charge density is then

$$\sigma = \frac{Q}{4\pi r^2} = 8.4 \times 10^{5}\,\mathrm{C/m}^2,$$

The repulsive force at a distance $r = 1$ m between the two balls is

$$F_C = \frac{1}{4\pi\varepsilon_0}\frac{Q^2}{r^2} = 1.56 \times 10^{15}\,\mathrm{N}.$$

and the electric field strength on the surface of each ball is

$$E = \frac{Q}{4\pi\varepsilon_0 r^2} = 9.6 \times 10^{16}\,\mathrm{V/m}.$$

1.2 (a) The total force must be directed into the direction of the string.

$$\Rightarrow \tan(\varphi/2) = \frac{F_{\mathrm{el}}}{m \cdot g}.$$

$$F_{\mathrm{el}} = \frac{Q^2}{4\pi\varepsilon_0 (2L \cdot \sin\varphi/2)^2}$$

$$\Rightarrow \frac{\sin^3(\varphi/2)}{\cos(\varphi/2)} = \frac{Q^2}{16\pi\varepsilon_0 L^2 \cdot mg}$$

Numerical values: $Q = 10^{-8}$ C, $m = 10$ g, $L = 1$ m

$$\Rightarrow \frac{\sin^3(\varphi/2)}{\cos(\varphi/2)} = \frac{10^{-16}}{16\pi\varepsilon_0 \times 10^{-2} \cdot 9.81} = 2.3 \times 10^{-6}$$

$$\Rightarrow \sin\varphi \approx \varphi, \cos\varphi \approx 1$$

$$\Rightarrow \varphi \approx 2 \cdot \sqrt[3]{2.3 \times 10^{-6}} = 2.6 \times 10^{-2}\mathrm{rad} \approx 1.5^{\circ}$$

$$\Rightarrow \text{Distance } r = 2L \cdot \sin\varphi/2$$

$$= 0.026\,\mathrm{m} = 2.6\,\mathrm{cm}.$$

$\Rightarrow$ the distance between the balls is $r = 2L \cdot \sin(\varphi/2)$.

(b) The conductive plate in the middle plane generates the electric field

$$\boldsymbol{E} = \frac{\sigma}{2\varepsilon_0}\hat{x} \Rightarrow \boldsymbol{F}_{\mathrm{el}} = \frac{Q \cdot \sigma}{2\varepsilon_0}\hat{x}$$

$$\Rightarrow \tan\varphi = \frac{F_{\mathrm{el}}}{m \cdot g} = \frac{Q \cdot \sigma}{2\varepsilon_0 m \cdot g}.$$

Numerical values: $Q = 10^{-8}$ C, $\sigma = 1.5 \times 10^{-5}\,\mathrm{C/m}^2$, $m = 0.05\,\mathrm{kg}$

$$\Rightarrow \tan\varphi = \frac{10^{-8} \cdot 1.5 \times 10^{-5}}{2 \cdot 8.85 \times 10^{-12} \cdot 0.05 \cdot 9.81}$$

$$= 1.7 \times 10^{-2}$$

$$\Rightarrow \varphi = 1^{\circ}.$$

$\Rightarrow$ Distance from the plate: $x = l \cdot \varphi = 17\,\mathrm{mm}$.

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4

1.3 (a) The force is, according to Fig. 1.11:

$$F = \int_{\alpha=\alpha_i}^{\alpha_a} \frac{q\cdot\sigma}{2\varepsilon_0}\sin\alpha\cdot d\alpha$$
$$= \frac{q\cdot\sigma}{2\varepsilon_0}(\cos\alpha_i - \cos\alpha_a),$$
$$\cos\alpha = \frac{x}{\sqrt{r^2+x^2}}$$
$$\Rightarrow F = \frac{q\sigma x}{2\varepsilon_0}\left[\frac{1}{\sqrt{R_i^2+x^2}} - \frac{1}{\sqrt{R_a^2+x^2}}\right],$$

(b) (α) $R_i \to 0$

$$F = \frac{q\cdot\sigma}{2\varepsilon_0}\left[1 - \frac{x}{\sqrt{R_a^2+x^2}}\right],$$

(β) $R_a \to \infty$:

$$F = \frac{q\cdot\sigma}{2\varepsilon_0}\frac{1}{\sqrt{1+R_i^2/x^2}},$$

(γ) $R_i \to 0$, $R_a \to \infty$

$$F = \frac{q\cdot\sigma}{2\varepsilon_0}.$$

1.4 The potential $\phi(r)$ is

$$\phi(r) = \frac{Q}{4\pi\varepsilon_0 R}.$$

Since the two balls are connected by conductors their potentials must be equal.

$$\Rightarrow \phi_1(R_1) = \frac{Q_1}{4\pi\varepsilon_0 R_1} = \phi_2(R_2) = \frac{Q_2}{4\pi\varepsilon_0 R_2}$$
$$\Rightarrow \frac{Q_1}{Q_2} = \frac{R_1}{R_2}, \quad Q = Q_1 + Q_2 = Q_1\left(1 + \frac{R_2}{R_1}\right)$$
$$\Rightarrow Q_1 = \frac{Q\cdot R_1}{R_1+R_2}, Q_2 = \frac{R_2\cdot Q}{R_1+R_2}$$
$$E_1 = \frac{Q_1}{4\pi\varepsilon_0 R_1^2} = \frac{Q}{4\pi\varepsilon_0 R_1(R_1+R_2)}$$
$$E_2 = \frac{Q_2}{4\pi\varepsilon_0 R_2^2} = \frac{Q}{4\pi\varepsilon_0 R_2(R_1+R_2)}.$$

1.5 According to Fig. A.1 is

(a) $Q_1 = Q_2 = Q$

$$\phi(R) = \frac{Q}{4\pi\varepsilon_0}\left(\frac{1}{r_1} + \frac{1}{r_2}\right).$$

For $R \gg a$ is

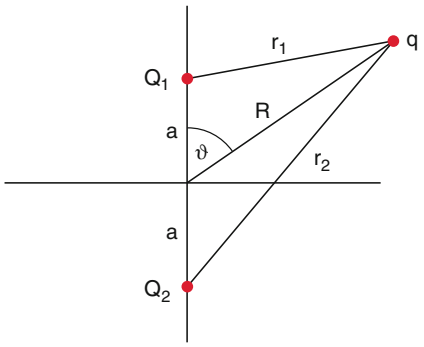


Fig. A.1 Illustration of problem 1.5

$$\phi(R) \approx \frac{Q}{4\pi\varepsilon_0}\left(\frac{1}{R - a\cos\vartheta} + \frac{1}{R + a\cos\vartheta}\right) \quad \text{for} \quad R \gg a$$
$$= \frac{2Q}{4\pi\varepsilon_0 R}\cdot\frac{1}{1 - \frac{a^2}{R^2}\cos^2\vartheta}.$$

The Taylor expansion of the fraction yields with

$$\frac{1}{1-x} \approx 1 + x + x^2 + \cdots + x^n$$
$$\Rightarrow \phi(R) = \frac{2Q}{4\pi\varepsilon_0 R}\left(1 + \frac{a^2}{R^2}\cos^2\vartheta + \frac{a^4}{R^4}\cos^4\vartheta + \cdots\right).$$

The force onto the charge q is obtained from

$$\boldsymbol{F} = -q\cdot\text{grad}\ \phi(r).$$

(b) $Q_1 = -Q_2 = Q$:

$$\phi = \frac{Q}{4\pi\varepsilon_0}\left(\frac{1}{r_1} - \frac{1}{r_2}\right)$$
$$= \frac{2aQ}{4\pi\varepsilon_0}\frac{\cos\vartheta}{R^2 - a^2\cos^2\vartheta}$$
$$= \frac{2|p|.\cos\vartheta}{4\pi\varepsilon_0 R^2}\frac{1}{1 - \frac{a^2}{R^2}\cos^2\vartheta}$$
$$= \frac{2p.\cos\vartheta}{4\pi\varepsilon_0 R^2}\left[1 + \frac{a^2}{R^2}\cos^2\vartheta + \frac{a^4}{R^4}\cos^4\vartheta + \cdots\right].$$

The comparison of (a) and (b) shows that for equal charges the first term gives the Coulomb potential of the total charge. This term is missing in (b) because $Q_1 + Q_2 = 0$. In (b) the series begins with the dipole term. For $R \gg a$ this term gives the major part of the force onto q.
(a) for $Q_1 = Q_2$:

$$\boldsymbol{F} = \frac{2Qq}{4\pi\varepsilon_0 R^2}\hat{\boldsymbol{R}},$$

(b) for $Q_1 = -Q_2$ see Eq. (1.25b).

1.6 (a)

$$E_{\text{pot}} = +3\frac{Q^2}{4\pi\varepsilon_0 a}$$

(b)

$$E_{\text{pot}} = \frac{1}{4\pi\varepsilon_0 a}\left(Q^2 - 2Q^2\right) = -\frac{Q^2}{4\pi\varepsilon_0 a}$$

(c)

$$E_{\text{pot}} = \frac{-4Q^2}{4\pi\varepsilon_0 a} + 2\frac{Q^2}{4\pi\varepsilon_0 a\sqrt{2}} = \frac{Q^2}{4\pi\varepsilon_0 a}\left(-4+\sqrt{2}\right) \approx -\frac{2.6Q^2}{4\pi\varepsilon_0 a}$$

1.7 We place the four charges in the x-y plane. For case (a) we get the positions:

$$Q_1 = +Q\colon \left(\frac{-a}{2}, 0\right);$$
$$Q_2 = +Q\colon \left(\frac{+a}{2}, 0\right);$$
$$r_1^2 = r_2^2 = \frac{a^2}{4};$$
$$Q_3 = -Q\colon \left(0, \frac{a}{2}\sqrt{3}\right);$$
$$Q_4 = -Q\colon \left(0, \frac{-a}{2}\sqrt{3}\right);$$
$$r_3^2 = r_4^2 = \frac{3a^2}{4}.$$

From the definition (1.36) we obtain:

$$QM_{xx} = Q_1\left(\frac{3}{4}a^2 - \frac{1}{4}a^2\right) + Q_2\left(\frac{3}{4}a^2 - \frac{1}{4}a^2\right) + Q_3\left(-\frac{3}{4}a^2\right) + Q_4\left(-\frac{3}{4}a^2\right) = \frac{5}{2}Qa^2,$$
$$QM_{yy} = -\frac{7}{2}Qa^2; \quad QM_{zz} = +a^2Q;$$
$$QM_{xy} = QM_{xz} = QM_{yz} = 0.$$

For case (b) we get:

$$Q_1 = -Q\colon (-a, 0); \quad Q_2 = 2Q : (0, 0);$$
$$Q_3 = -Q\colon (+a, 0).$$

this yields with (1.36) the result:

$$QM_{xx} = -4Qa^2; \quad QM_{yy} = 2Qa^2;$$
$$QM_{zz} = 2Qa^2;$$
$$QM_{xy} = QM_{xz} = QM_{yz} = 0.$$

1.8 Analogue to the calculation of the gravitational potential in Vol. 1, Sect. 2.9.5 the electrical potential $\phi(r)$ at the point $P(r)$ of a homogeneously charged ball with radius R

(a) For $r \le R$:

$$\phi(r) = \frac{Q}{8\pi\varepsilon_0 R^3}\left(3R^2 - r^2\right) \quad \text{with } Q = \frac{4}{3}\pi\varrho_{\text{el}}R^3$$
$$= \frac{1}{6}\frac{\varrho_{\text{el}}}{\varepsilon_0}\left(3R^2 - r^2\right).$$

(b) For $r \ge R$:

$$\phi(r) = \frac{Q}{4\pi\varepsilon_0 r}.$$

The work necessary to bring a charge q from $r = 0$ to $r = R$ is

$$W_1 = q\cdot[\phi(R) - \phi(0)] = \frac{q\cdot Q}{4\pi\varepsilon_0 R}\cdot\left(1 - \frac{3}{2}\right) = -\frac{q\cdot Q}{8\pi\varepsilon_0 R}.$$

On the way from $r = R$ to $r = \infty$ the work

$$W_2 = -\frac{q\cdot Q}{4\pi\varepsilon_0 R},$$

that has to be spend is twice that of W_1. The electric field strength is

$$E(r) = -\frac{\mathrm{d}\phi(r)}{\mathrm{d}r}:$$

$$\boldsymbol{E}(r) = \frac{Q\cdot r}{4\pi\varepsilon_0 R^3}\hat{\boldsymbol{r}} \quad \text{for} \quad r \le R,$$

$$\boldsymbol{E}(r) = \frac{Q}{4\pi\varepsilon_0 r^2}\hat{\boldsymbol{r}} \quad \text{for} \quad r \ge R.$$

1.9 The explicit calculation of the Taylor expansion is as follows:

$$\frac{1}{|\boldsymbol{R}-\boldsymbol{r}|}=\frac{1}{R}-\left(x\frac{\partial}{\partial X}\frac{1}{R}+y\frac{\partial}{\partial Y}\frac{1}{R}+z\frac{\partial}{\partial Z}\frac{1}{R}\right)$$
$$+\frac{1}{2}\left[xx\frac{\partial^2}{\partial X^2}\frac{1}{R}+xy\frac{\partial^2}{\partial X\partial Y}\frac{1}{R}+\cdots+zz\frac{\partial^2}{\partial Z^2}\frac{1}{R}\right]+\cdots.$$
$$R=\sqrt{X^2+Y^2+Z^2}\Rightarrow\frac{\partial}{\partial X}\frac{1}{R}=\frac{-X}{R^3}$$
$$\frac{\partial^2}{\partial X\partial Y}\frac{1}{R}=\frac{3}{2}\frac{X\cdot Y}{R^5}\quad\text{etc.}$$

Corresponding expressions are obtained for the other derivations. Inserting these expressions the potential becomes:

$$\phi(R)=\frac{1}{4\pi\varepsilon_0}\sum\frac{Q_i}{|\boldsymbol{R}-\boldsymbol{r}_i|}$$
$$=\frac{1}{4\pi\varepsilon_0}\left[\frac{\sum Q_i}{R}+\frac{1}{R^3}\sum(Q_i\boldsymbol{r}_i)\cdot\boldsymbol{R}+\frac{1}{R^5}\frac{1}{2}\sum Q_i\{(3x_i^2-r_i^2)X^2\right.$$
$$+(3y_i^2-r_i^2)Y^2+(3z_i^2-r_i^2)Z^2+2\cdot3x_iy_iXY$$
$$\left.+2\cdot3y_iz_iYZ+2\cdot3x_iz_iXZ\}\right].$$

1.10 In order to prove that only the monopole term does not vanish one must show that

$$\phi(R)=\frac{1}{4\pi\varepsilon_0}\int_V\frac{\varrho_{\mathrm{el}}}{|\boldsymbol{R}-\boldsymbol{r}|}\mathrm{d}V=\frac{Q}{4\pi\varepsilon_0R}.$$

The charge $\mathrm{d}Q=\rho_{\mathrm{el}}\cdot2\pi y\cdot\mathrm{d}y\cdot\mathrm{d}x$ on the circular ring with radius y and a distance x of the ring plane from the center $x=y=0$ (see Fig. A.2) have the same distance

$$r=\sqrt{y^2+(R-x)^2}$$

from the point $P(R)$ and contribute the share

$$\mathrm{d}\phi=\frac{\mathrm{d}Q}{4\pi\varepsilon_0r}=\frac{\varrho_{\mathrm{el}}}{2\varepsilon_0}\frac{y\,\mathrm{d}y}{\sqrt{y^2+(R-x)^2}}\mathrm{d}x.$$

to the potential in $P(R)$.

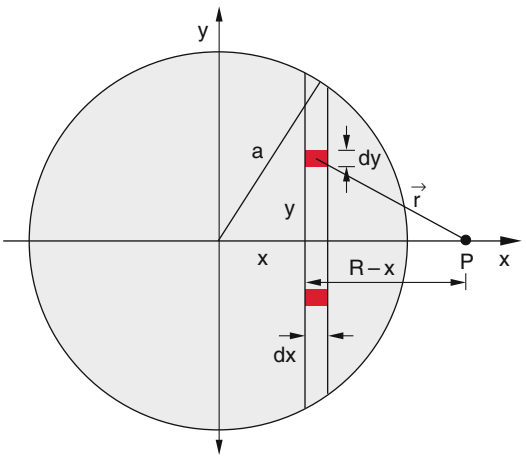


Fig. A.2 Illustration of problem 1.10

The potential generates by the whole circular disc is then

$$\phi_{\mathrm{dis}}=\frac{\varrho_{\mathrm{el}}}{2\varepsilon_0}\left[\int_{y=0}^{\sqrt{a^2-x^2}}\frac{y\,\mathrm{d}y}{\sqrt{y^2-(R-x)^2}}\mathrm{d}x\right].$$

Integration from $x=-a$ to $x=+a$ yields the contribution of the sphere

$$\phi_{\mathrm{sphere}}=\frac{\varrho_{\mathrm{el}}}{\varepsilon_0}\frac{a^3}{3R}=\frac{Q}{4\pi\varepsilon_0R}.$$

1.11 The electric field strength E of a single wire is according to (1.18a) for $r\geq R$

$$\boldsymbol{E}=\frac{\lambda}{2\pi\varepsilon_0r}\hat{\boldsymbol{r}}$$

with $\lambda=Q/L=$ charge per unit length of the wire. The total field strength of the charge distribution in Fig. A.3 is along the x-axis for $|x|<a$

$$\boldsymbol{E}=\{E_x,0,0\}$$
$$E_x=\frac{\lambda}{2\pi\varepsilon_0}\left[\frac{1}{a+x}-\frac{1}{a-x}+\frac{2x}{a^2+x^2}\right]$$
$$=\frac{\lambda}{2\pi\varepsilon_0}\left[\frac{-2x}{a^2-x^2}+\frac{2x}{a^2+x^2}\right]$$
$$=-\frac{\lambda}{\pi\varepsilon_0}\frac{2x^3}{a^4-x^4}.$$

For $x=0$ is $E=0$, For $x=a-R$ (i.e. on the inner surface of the wire) is

$$E_x=-\frac{\lambda}{\pi\varepsilon_0}\frac{2(a-R)^3}{a^4-(a-R)^4}$$
$$=-\frac{\lambda}{2\pi\varepsilon_0R}\cdot\frac{4\cdot R(a-R)^3}{a^4-(a-R)^4}.$$

With $a=4$ cm and $R=0.5$ cm this becomes

$$E_x=-\frac{\lambda}{2\pi\varepsilon_0R}\cdot0.8\,\mathrm{V/m}.$$

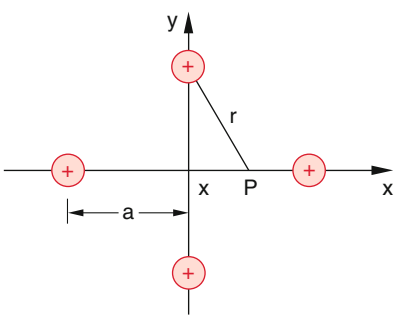


Fig. A.3 Illustration of problem 1.11

The electric field strength at the surface at ($x = a - R$) is only 80% of the field of a single wire with equal charge density. For the outer surface ($x = a + R$) the electric field

$$E = \frac{\lambda}{2\pi\varepsilon_0 R} \cdot \frac{4R(a+R)^3}{(a+R)^4 - a^4} = \frac{\lambda}{2\pi\varepsilon_0 R} \cdot 1.18\,\mathrm{V/m}$$

is slightly larger than that of a single wire. However, one can transport the fourfold electric power 4P. If the single wire should transport the same power $P = U^2/R$ one has to double the voltage U.

1.12 (a) One obtains

$$\begin{aligned} C &= \varepsilon_0 \cdot \frac{A}{d} = \frac{8.85 \times 10^{-12} \cdot 0.1}{0.01}\,\mathrm{F} \\ &= 8.85 \times 10^{-11}\ \mathrm{F} = 88.5\,\mathrm{pF}; \\ Q &= C \cdot U = 8.85 \times 10^{-11} \cdot 5 \times 10^3\,\mathrm{C} \\ &= 4.4 \times 10^{-7}\,\mathrm{C}; \\ E &= \frac{U}{d} = 5 \times 10^5\,\mathrm{V/m}. \end{aligned}$$

(b) If the capacitor with the voltage U_0 is discharged through the resistor R the total energy $W = C \cdot U^2$ is converted into Joule's heat energy. We then get

$$W = \int_0^\infty I^2 \cdot R \cdot \mathrm{d}t.$$

With $I = (U_0/R) \cdot \mathrm{e}^{-t/(RC)}$ (see (2.10)) this gives

$$\begin{aligned} W &= \frac{U_0^2}{R} \cdot \left(-\frac{R \cdot C}{2}\right) \cdot \mathrm{e}^{-2t/(RC)}\Big|_0^\infty \\ &= \frac{U_0^2 C}{2}. \end{aligned}$$

(c) $\boldsymbol{D} = \boldsymbol{p} \times \boldsymbol{E}$

$$\begin{aligned} \Rightarrow |\boldsymbol{D}| &= 1.6 \times 10^{-19} \cdot 5 \times 10^{-11} \cdot 5 \times 10^5\,\mathrm{N\,m} \\ &= 4 \times 10^{-24}\,\mathrm{N\,m}, \\ W_{\mathrm{pot}} &= p \cdot \boldsymbol{E} = 4 \times 10^{-24}\,\mathrm{N\,m}. \end{aligned}$$

1.13 Figure 1.69 can be redesigned into the equivalent circuit in Fig. A.4. Then we get

$$\frac{1}{C_\mathrm{g}} = \frac{1}{C} + \frac{1}{3C} \Rightarrow C_\mathrm{g} = \frac{3}{4}C.$$

The capacity in the dotted box is

$$C + \frac{1}{2}C = \frac{3}{2}C.$$

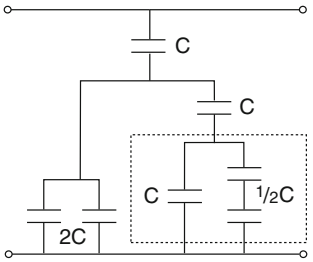


Fig. A.4 Illustration of problem 1.13

We then get for the right part in Fig. A.4

$$\frac{1}{C} + \frac{1}{3/2C} = \frac{5}{3C} \Rightarrow C_r = \frac{3}{5}C.$$

For the right and left part is

$$\frac{3}{5}C + 2C = \frac{13}{5}C$$

The total capacity is then

$$C_\mathrm{total} = (13/18)C$$

1.14 On the right plate in Fig. 1.70 the charge $-Q/2$ is accumulated by influence. This charge must be taken from the left plate of the right capacitor, where the residual charge $+Q/2$ remains. We therefore get electric field and the potential as shown in (Fig. A.4)

$$E = \frac{U}{d} = \frac{3}{4}\frac{Q}{C \cdot d}.$$

1.15 (a) For the calculation of the potential ϕ we write the Laplace equation (1.16b) in cylindrical coordinates (note, that ϕ does not depend on z and φ).

$$\begin{aligned} \Delta\phi &= \frac{1}{R}\frac{\partial}{\partial R}\left(R \cdot \frac{\partial \phi}{\partial R}\right) = 0 \qquad (1) \\ \Rightarrow \phi &= c_1 \mathrm{In\,R} + c_2 \end{aligned}$$

With $\phi(R_1) = \phi_1$, $\phi(R_2) = \phi_2$ it follows

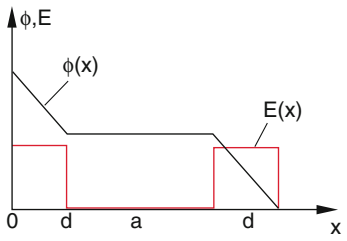


Fig. A.5 Illustration of problem 1.14

$$c_2 = \phi_1 - c_1 \ln R_1,$$
$$c_1 = \frac{\phi_2 - \phi_1}{\ln(R_2/R_1)} \qquad (2)$$
$$\Rightarrow \phi(R) = \phi_1 + \frac{\phi_2 - \phi_1}{\ln(R_2/R_2)} \ln(R/R_1),$$

$$E(R) = \frac{\partial \phi}{\partial R} = \frac{-(\phi_2 - \phi_1)}{\ln(R_2/R_1)} \frac{1}{R}. \qquad (3)$$

For the desired circular path with radius $R_0 = (R_1 + R_2)/2$ the relation holds

$$\frac{m v_0^2}{R_0} = e \cdot E(R_0) = \frac{2e}{R_1 + R_2} \frac{\phi_1 - \phi_2}{\ln (R_2/R_1)}$$
$$\Rightarrow U = \frac{R_1 + R_2}{2e} \ln(R_2/R_1) \cdot \frac{m}{R} v_0^2$$
$$= \frac{R_1 + R_2}{2R} \frac{m}{e} v_0^2 \ln \frac{R_2}{R_1}.$$

For $R = (R_1 + R_2)/2$ is

$$U = \frac{m}{e} v_0^2 \ln \frac{R_2}{R_1}. \qquad (4)$$

(b) Assume an electron enters the cylindrical capacitor at $r = R_0$, $\varphi = 0$ with $|v| = |v_0|$, but under a small angle α against the desired path $R = R_0$. Can the electron intersect the desired circle $R = R_0$?. At which angle φ does this happen?
Since $E(r)$ is a central field the angular momentum of the particles remains constant, i.e.

$$v \cdot R = v_0 \cdot R_0 = \text{const.} \qquad (5)$$

When the deviation at time t is δR the equation of motion is

$$m \cdot \delta \ddot{R} - m \cdot \frac{v^2}{R} - e \cdot E(R_0 + \delta R) = 0. \qquad (6)$$

Expansion into a Taylor series yields

$$E(R_0 + \delta R) = E(R_0) + \left(\frac{dE}{dR}\right)_{R_0} \delta R + \cdots. \qquad (7)$$

From (6) it follows

$$\frac{dE}{dR} = \frac{U}{\ln(R_2/R_1)} \frac{1}{R^2}.$$

Inserting into (9) gives with (8)

$$\delta \ddot{R} - \frac{v_0^2}{R^3} R_0^2 + \frac{v_0^2}{R_0} \left(1 - \frac{\delta R}{R_0}\right) = 0.$$
$$\frac{1}{R^3} = \frac{1}{R_0^3 \left(1 + \frac{\delta R}{R_0}\right)^3} \approx \frac{1}{R_0^3} - \frac{3}{R_0^4} \delta R + \cdots$$
$$\Rightarrow \delta \ddot{R} - \frac{v_0^2}{R_0} \left(1 - 3\frac{\delta R}{R_0} - 1 + \frac{\delta R}{R_0}\right) = 0$$
$$\Rightarrow \delta \ddot{R} + 2\omega_0^2 \delta R = 0 \quad \text{with} \quad \omega_0 = \frac{v_0}{R_0}.$$

The motion proceeds on a circular path with superimposed radial oscillation.

$$\delta R = R_0 \cdot \sin\left[\sqrt{2}\omega_0 \cdot t\right],$$

which becomes zero after the time $t = \pi/(\sqrt{2}\omega_0) \Rightarrow \varphi = \pi/\sqrt{2} = 127°$.

A cylindrical capacitor with $\varphi = 127°$ focusses the divergent incident particles.

1.16 The charge density of the wire is $\lambda = Q/L$. The element dL produces at $P = (0, 0)$ the electric field

$$d\boldsymbol{E} = \frac{1}{4\pi\varepsilon_0} \frac{\lambda \cdot d\boldsymbol{L}}{R^2} \{\cos \varphi, \sin \varphi, 0\}$$

The field generated by the total charged wire is (see Fig. A.6)

$$E_x = \frac{1}{4\pi\varepsilon_0} \frac{\lambda}{R^2} \int_{\varphi_1}^{\varphi_2} R \cdot \cos \varphi \, d\varphi,$$

$$E_y = \frac{1}{4\pi\varepsilon_0} \frac{\lambda}{R^2} \int_{\varphi_1}^{\varphi_2} R \cdot \sin \varphi \, d\varphi,$$

$$\varphi_1 = \frac{\pi}{2} - \frac{\alpha}{2} = \frac{\pi}{2} - \frac{L}{2R},$$
$$\varphi_2 = \frac{\pi}{2} + \frac{L}{2R}$$
$$\Rightarrow E_x = \frac{1}{4\pi\varepsilon_0} \frac{\lambda}{R} (\sin \varphi_2 - \sin \varphi_1)$$
$$= \frac{1}{4\pi\varepsilon_0} \frac{\lambda}{R} \left(\cos\frac{L}{2R} - \cos\frac{L}{2R}\right) = 0,$$
$$E_y = \frac{1}{4\pi\varepsilon_0} \frac{\lambda}{R} (\cos \varphi_1 - \cos \varphi_2)$$
$$= \frac{1}{4\pi\varepsilon_0} \frac{2\lambda}{R} \sin\frac{L}{2R}.$$

The field E has therefore only a y-component and its amount is

$$|\boldsymbol{E}| = \frac{1}{2\pi\varepsilon_0} \frac{\lambda}{R} \sin\frac{L}{2R}.$$

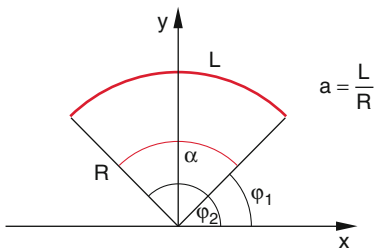


Fig. A.6 Illustration of problem 1.16

Chapter 2

2.1 (a) The mass of a Cu-atom is 63.5×10^{-27} kg. The number of atoms per m^3 is

$$n = \frac{8.92 \times 10^3}{63.5 \cdot 1.66 \times 10^{-27}\,\mathrm{m}^{-3}} = 8.5 \times 10^{28}/\mathrm{m}^3$$

$\Longrightarrow$ on the average one free electron is present per 8.5/5 = 1.7 atoms.

(b) The electric current flows through the light bulb after a time

$$t_1 = \frac{L}{c} = \frac{10\,\mathrm{m}}{3 \times 10^8\,\mathrm{m/s}} \approx 3 \times 10^{-8}\,\mathrm{s},$$

i.e. practically instantaneously. Because the filament of the bulb heats up, its electrical resistance increases from R_0 to R. The current therefore decreases from the initial value $I_0 = U/R_0$ to $I = U/R = P_{el}/U$ when P_{el} is the electrical power of the light bulb. The temperature of the filament increases up to a value T_m where the power radiated by the hot filament equals the electrical power supplied to the light bulb.

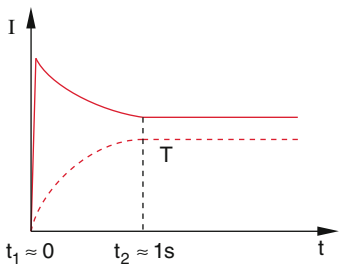


Fig. A.7 To solution of problem 2.1b

(c) The current density is

$$j = \frac{I}{\pi r^2} = 2.6 \times 10^6\,\mathrm{A/m}^2.$$

For $j = e \cdot n \cdot v_D$ follows with $n = 5 \times 10^{28}/\mathrm{m}^3$ the drift velocity $v_D = 0.33 \times 10^{-3}$ m/s = 0.33 mm/s $\Longrightarrow$ $t_2 = 3 \times 10^4$ s.

It takes about 8 h (!) until the first electron from the current source reaches the filament.

(d) At a current of 1 A pass $N = 6.25 \times 10^{18}$ electrons per sec through the cross section of the wire. Their total mass is

$$M = 6.25 \times 10^{18} \cdot 9.1 \times 10^{-31}\,\mathrm{kg} = 5.6 \times 10^{-12}\,\mathrm{kg}.$$

It takes therefore 1.7×10^{11} s = 5.4 years (!) until 1 g of electrons has passed through the filament.

2.2 The electrical resistance dR of the conductor element dx is

$$\mathrm{d}R = \varrho_{el} \cdot \frac{\mathrm{d}x}{A(x)}.$$

The cross section of the wire is

$$A(x) = \frac{\pi}{4}(\mathrm{d}(x))^2 = \frac{\pi}{4}\left(d_1 + \frac{d_2 - d_1}{L}x\right)^2.$$

The total resistance is then

$$R = \frac{4\varrho_{el}}{\pi}\int_0^L \left(d_1 + \frac{d_2 - d_1}{L}x\right)^{-2}\mathrm{d}x$$

$$= \frac{4\varrho_{el}}{\pi}\int_0^L \frac{\mathrm{d}x}{(a+bx)^2}$$

with $a = d_1,\ b = (d_2 - d_1)/L$

$$= \frac{4\varrho_{el}}{\pi \cdot b}\frac{1}{(a+bx)}\Big|_0^L = \frac{4\varrho_{el}}{\pi} \cdot \frac{L}{d_1 \cdot d_2}.$$

Numerical values:

$$R = \frac{4 \cdot 8.71 \times 10^{-8}}{\pi} \cdot \frac{1}{0.25 \times 10^{-6}\,\Omega} = 0.44\,\Omega.$$

(b) At the voltage $U = 1$ V the current is

$$I = \frac{1}{0.44}\,\mathrm{A} \approx 2.25\,\mathrm{A}.$$

For the total length of the wire the electric power consumption is $P_{el} = U \cdot I = 2.25$ W. It is not uniformly distributed along the wire because of the changing cross section of the wire. With $\mathrm{d}P_{el} = I^2 \cdot \mathrm{d}R$ we get

$$P_{el}(x) = I^2 \cdot \varrho_{el} \cdot \frac{\mathrm{d}x}{A(x)}.$$

The electric power consumption is inversely proportional to the cross section A(x) of the wire (Fig. A.7).

2.3 The two resistors $2R$ in the middle of Fig. 2.75 are shortened and should therefore not be taken into account. Between B and the middle point M the total resistance is $4R/5$. The same holds for the resistance between A and M. The total resistance R_t between A and B is then $R_t = 8R/5$.

2.4 The circuit in Fig. 2.76 can be arranged in a simplified form as shown in Fig. A.8:

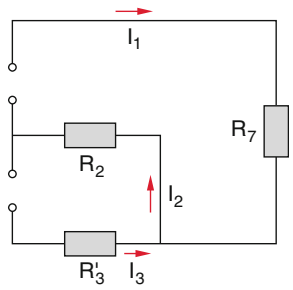


Fig. A.8 Illustration of problem 2.4

$$R_3' = R_3 + R_i(U_2) = (4+1)\,\Omega = 5\,\Omega$$
$$R_7 = R_1 + R_i(U_1) + R_4 + \frac{R_5 \cdot R_6}{R_5 + R_6}$$
$$= \left(3 + 1 + 8 + \frac{12.24}{36}\right)\Omega = 20\,\Omega$$

(a) $I_1 + I_3 = I_2$ (Kirchhoff's current law)
(b) $I_1 \cdot R_7 + I_2 \cdot R_2 = U_1$ (upper mesh)
(c) $I_3 \cdot R_3' + I_2 \cdot R_2 = U_2$ (lower mesh)

From (b) it follows $I_1 = \dfrac{U_1 - I_2 R_2}{R_7}$.

From (c) it follows $I_3 = \dfrac{U_2 - I_2 R_2}{R_3'}$.

Inserting into (a) gives for I_2, I_1 and I_3:

$$I_2 = \frac{U_1 R_3' + U_2 R_7}{R_2\left(R_3' + R_7\right) + R_3' R_7} = 0.65\,\text{A}.$$
$$I_1 = \frac{U_1}{R_7} - \frac{R_2}{R_7} I_2 = 0.37\,\text{A};$$
$$I_3 = I_2 - I_1 = 0.28\,\text{A}.$$

The potential difference is equal to the voltage

$$U(A) = \frac{R_5 \cdot R_6}{R_5 + R_6} \cdot I_1 = 2.96\,\text{V}.$$

2.5 (a) $U_1 = U_0 - IR_i$

$$\Rightarrow R_i = \frac{U_0 - U}{I} = \frac{2}{150}\,\Omega = 13.3\,\text{m}\Omega$$
$$R_\text{a} = \frac{U_1}{I} = \frac{10}{150}\Omega = 66.7\,\text{m}\Omega.$$

(b) For $R_\text{i} = R_\text{a}$ is

$$I = \frac{U_1}{R_\text{a}} = \frac{U_0 - IR_\text{a}}{R_\text{a}}$$
$$\Rightarrow I = \frac{U_0}{2R_\text{a}} = \frac{12}{0.133}\,\text{A} = 90\,\text{A}$$
$$U_1 = U_0 - IR_\text{a}$$
$$= (12 - 90 \cdot 0.0667)\text{V} = 6\text{V}.$$

(c) For the case (a) the power, consumed in the starter is

$$P_\text{el}^\text{A} = I^2 \cdot R_a = 150^2 \cdot 0.0667\,\text{W} = 1500\,\text{W}$$

In the battery is the power, consumed during the starter process

$$P_\text{el}^{(\text{B})} = I^2 \cdot R_i = 150^2 \cdot 0.0133\,\text{W} \approx 300\,\text{W}$$

For case (b) is

$$P_\text{el}^{(\text{A})} = 90^2 \cdot 0.0667\,\text{W} \approx 540\,\text{W}$$
$$P_\text{el}^{(\text{B})} = 540\,\text{W}$$

2.6 We conflate the elements 1–8 as follows:

Zusammenfassung	Art	C_g	R_g
$7 + 8 = a$	Serie	$\frac{1}{2}C$	$2R$
$6 + a = b$	Parallel	$\frac{3}{2}C$	$\frac{2}{3}R$
$5 + b = c$	Serie	$\frac{3}{5}C$	$\frac{5}{3}R$
$4 + c = d$	Parallel	$\frac{8}{5}C$	$\frac{5}{8}R$
$3 + d = e$	Serie	$\frac{8}{13}C$	$\frac{13}{8}R$
$2 + e = f$	Parallel	$\frac{21}{13}C$	$\frac{13}{21}R$
$1 + f$	Serie	$\frac{21}{34}C$	$\frac{34}{21}R$

$$\Rightarrow C_g = \frac{21}{34}C, R_g = \frac{34}{21}R.$$

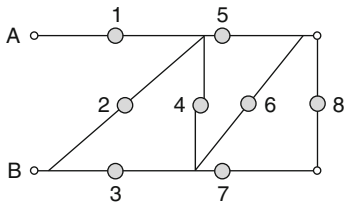


Fig. A.9 Illustration of solution 2.6

2.7 The wanted nickel layer with thickness d has the volume

$$V = d \cdot A = d\left(2\pi r \cdot L + 2\pi r^2\right) = 24.9\,\text{cm}^3$$

Its mass is

$$m = \varrho \cdot V = 8.7 \cdot 24.9\,\text{g} = 216.5\,\text{g}.$$

(a) The total current I is equal to the acceptable current density j times the surface A of the cylinder.

$$I = 2.5 \times 10^{-1}\,\text{A/cm}^2 \cdot 2.49 \times 10^3\,\text{cm}^2 = 623\,\text{A}.$$

(b) The electro-chemical equivalent is

$$E_C = \frac{1}{2} \cdot \frac{N_A \cdot m_{Ni}}{96485.3} \frac{\text{kg}}{\text{C}} = 1.825 \times 10^{-7}\,\text{kg/C}$$
$$= 1.825 \times 10^{-4}\,\text{g/C}.$$

The total time for galvanizing the cylinder is

$$t = \frac{216.5}{1.825 \times 10^{-4} \cdot 623}\,\text{s} = 1.9 \times 10^3\,\text{s} = 31.7\,\text{min}.$$

2.8 With the open circuit voltage U_0, the voltage U under workload is

$$U = U_0 - I \cdot R_i$$

$$I = \frac{U}{R_a} \Rightarrow U = \frac{U_0}{1 + R_i/R_a}$$

$$P_{el} = \frac{dW_{el}}{dt} = \frac{U^2}{R_a} = \frac{U_0^2 R_a}{(R_i + R_a)^2}$$

$$\frac{dP_{el}}{dR_i} = 0 \Rightarrow R_i = R_a$$

$$\Rightarrow P_{el}^{max} = \frac{U_0^2}{4R_i} = \frac{4.5}{4 \cdot 1.2}\,\text{W} = 4.22\,\text{W}.$$

2.9 (a)

$$Q = C_1 U_1 = 2 \cdot 10^{-5}\text{F} \times 10^3\,\text{V} = 2 \times 10^{-2}\,\text{C}$$

After connecting the two capacitors with each other, the charge is distributed onto C_1 and C_2 in such a way that the voltage of the two capacitors has the same value U_2.

$$Q = (C_1 + C_2)U_2$$
$$\Rightarrow U_2 = \frac{Q}{C_1 + C_2}$$
$$= \frac{2 \times 10^{-2}\,\text{C}}{3 \times 10^{-5}\text{F}} = \frac{2}{3} \times 10^3\,\text{V}.$$

Before the connection the energy was

$$W_{el} = \frac{1}{2} C_1 U_1^2 = 10\,\text{Ws}.$$

After the connection is

$$W_1 = \frac{1}{2} C_1 U_2^2 = \frac{40}{9}\,\text{Ws}$$

$$W_2 = \frac{1}{2} C_2 U_2^2 = \frac{20}{9}\,\text{Ws}$$

$$\Rightarrow W = W_1 + W_2 = \frac{20}{3}\,\text{Ws}.$$

The difference ΔW = 10/3 Ws has been consumed as Joule' heat by the current from C_1 to C_2 during the recharging process.
This can be expressed also by

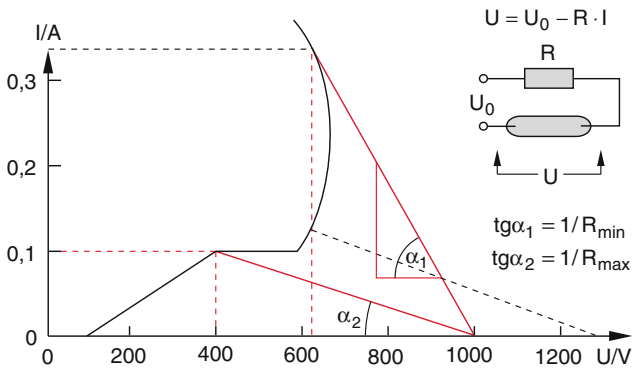


Fig. A.10 Illustration of solution 2.10

$$W_{el} = \frac{1}{2}\frac{Q^2}{C_1}, \quad W_1 + W_2 = \frac{1}{2}\frac{Q^2}{C_1 + C_2} < W_{el}$$

$\Longrightarrow$ the fraction $C_2/(C_1 + C_2)$ of the initial energy is lost as heat energy (Fig. A.9).

2.10 From Fig. A.10 we obtain

$$U = U_0 - R \cdot I.$$

A stable discharge is possible up to the turning point in Fig. A.10. This gives for the turning point the values U = 630 V and I = 0.33 A.
For R_{min} is the straight line $I(U) = (U_0 - U)/R$ the tangent to the current-voltage characteristic curve of the gas discharge. For U_0 = 1000 V and U = 630 V is then I = 0.33 A,

$$\Rightarrow R_{min} = \frac{U_0 - U}{I} = \frac{1000 - 630}{0.33}\Omega \approx 1121\,\Omega$$
$$R_{max} = \frac{1000 - 400}{0.1}\Omega = 6000\,\Omega.$$

(a) For R = 5 kΩ and U_0 = 500 V is

$$I = \frac{U_0 - U}{R} = 0.1\,\text{A} - \frac{U}{R}.$$

The intersection of the straight resistance line with the current-voltage characteristic curve is located in the dependent part. The discharge extinguishes.
For U_0 = 1250 V is

$$I = \frac{1240\,\text{V}}{5000\,\Omega} - \frac{U}{5000\,\Omega} = 0.25\,\text{A} - \frac{U}{5000\,\Omega}.$$

Since U < 700 V lies I in the range 0.25 A > I > 0.11 A. The straight resistance line intersects the current-voltage characteristic is the stable range. From the graphical representation we get U = 620 V, I = 0.12 A.

2.11 $j = (n^+ + n^-)e \cdot v = \sigma \cdot E$

$$E = E_0 \cdot \cos\,\omega t$$
$$v = \frac{\sigma}{(n^+ + n^-)e} E_0 \cos\omega t = v_0 \cdot \cos\omega t$$
$$v_0 = \frac{1.1 \cdot 3000}{2 \times 10^{28} \cdot 1.6 \times 10^{-18}} \frac{\mathrm{m}}{\mathrm{s}} = 1 \times 10^{-6}\,\mathrm{m/s}$$
$$s_0 = \frac{v_0}{\omega}, \quad \text{because} \quad s = \int v\,\mathrm{d}t = \frac{1}{\omega} v_0 \sin\omega t$$
$$s_0 = 3.2 \times 10^{-9}\,\mathrm{m} = 3.2\,\mathrm{nm}.$$

2.12 According to (2.15) we get with $h = L$

$$R = \frac{\varrho_s \cdot \ln(r_2/r_1)}{2\pi \cdot L}$$
$$= \frac{10^{12} \ln 8}{200\pi} = 3.3 \times 10^9\,\Omega,$$
$$I = \frac{U}{R} = \frac{3 \times 10^3}{3.3 \times 10^9}\,\mathrm{A} = 0.9 \times 10^{-6}\,\mathrm{A} = 0.9\,\mu\mathrm{A}.$$

2.13 The resistance for n m cable length is

$$R_n = 2R_1 + R_{n-1}$$

where

$$\frac{1}{R_{n-1}} = \frac{1}{R_2} + \frac{n-1}{2R_1 + R_2}$$
$$\Rightarrow R_{n-1} = \frac{R_2(R_1 + R_2)}{2R_1 + n \cdot R_2}.$$

(b) For $R_1 = R_2$:

$$\Rightarrow R_{n-1} = \frac{3R_1}{2+n} \quad \Rightarrow R_n = 2R_1 + \frac{3R_1}{2+n},$$
$$\Rightarrow \lim_{n\to\infty} R_n = 2R_1.$$

2.14 With the mean free path length Λ an electron at the position r has suffered its last collision at the position $r - \Lambda$ where its velocity was

$$\boldsymbol{v}(r - \Lambda) = \langle v \rangle (T(r - \Lambda))\bar{\boldsymbol{e}}$$

where $\hat{\mathbf{e}}$ is the unit vector in the direction of $\mathbf{v}$. The mean velocity <v> depends on the temperature T. The mean velocity at the position r is obtained by integration over all directions, because only the direction but not the amount of v is altered at each elastic collision. This gives:

$$\langle v \rangle = \frac{1}{4\pi} \int \hat{\boldsymbol{v}} \cdot \bar{v}\,(T(r - \Lambda\hat{\boldsymbol{v}}))\,\mathrm{d}\Omega$$

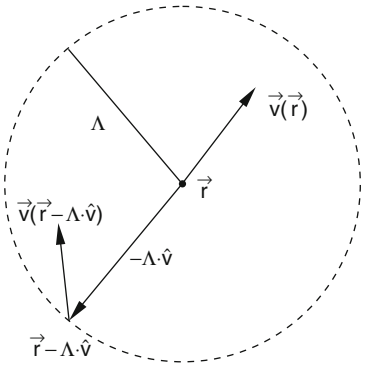


Fig. A.11 Illustration of solution 2.14

If the temperature changes only slightly over the mean path length Λ one needs only to keep the first two members in the Taylor expansion

$$\langle v \rangle T(r - \Lambda) = (T(r)) - \Lambda \cdot \nabla T(r) \cdot \mathrm{d}\bar{v}/\mathrm{d}T.$$

and can neglect all higher terms. This gives

$$\langle v \rangle T(r - \Lambda) \approx \langle v \rangle T(r) \cdot \Lambda \cdot \langle v \rangle \cdot \nabla T(r) \cdot \mathrm{d}\langle v \rangle / \mathrm{d}T.$$

Inserting this expression into the integral gives for the first term the value zero because the velocities are uniformly distributed over all directions. For the second term one obtains for the drift velocity

$$\boldsymbol{n}(r) = \langle v \rangle_r = -\frac{1}{4\pi} \nabla T(r) \cdot \frac{\mathrm{d}\bar{v}}{\mathrm{d}T} \cdot \int \Lambda\hat{v}\,\mathrm{d}\Omega$$
$$= -\frac{1}{3} \Lambda \cdot \frac{\mathrm{d}\bar{v}}{\mathrm{d}T} \cdot \nabla T(r).$$

The current density caused by thermo-diffusion is then

$$\boldsymbol{j}(r)_{TD} = \boldsymbol{n} \cdot \boldsymbol{v}(\boldsymbol{r}).$$

2.15 Without thermo-diffusion is $j_{\mathrm{TD}} = 0$. According to (2.42h) the *Seebeck* coefficient is then also zero. According to (2.41a) the thermo-voltage is determined by the difference of the *Seebeck* coefficients. Without thermo-diffusion therefore also the thermo-voltage is zero.

Chapter 3

3.1 (a) $B(0) = 0$: Outside the wires the fields add, inside they subtracted each other.

$$F_1 = \{+F_x, 0, 0\}, \quad F_2 = \{-F_x, 0, 0\}$$

(b) For the magnetic field we get:

$$|\boldsymbol{B}_1| = B_1 = \frac{\mu_0 I_1}{2\pi r_1},$$
$$B_{1x} = B_1 \cdot \sin\alpha_1 = B_1 \frac{a-y}{r_1},$$
$$B_{1y} = B_1 \cdot \cos\alpha_1 = B_1 \frac{x}{r_1},$$
$$|\boldsymbol{B}_2| = B_2 = \frac{\mu_0 I_2}{2\pi r_2},$$
$$B_{2x} = -B_2 \sin\alpha_2 = -B_2 \frac{a+y}{r_2},$$
$$B_{2y} = B_2 \cos\alpha_2 = B_2 \frac{x}{r_2}.$$

The total field at the point $P(x, y)$ is

$$B_x = \frac{a-y}{r_1} B_1 - \frac{a+y}{r^2} B_2$$
$$= \frac{\mu_0}{2\pi}\left(\frac{I_1(a-y)}{r_1^2} - \frac{I_2(a+y)}{r_2^2}\right),$$
$$B_y = \frac{\mu_0 x}{2\pi}\left(\frac{I_1}{r_1^2} + \frac{I_2}{r_2^2}\right)$$

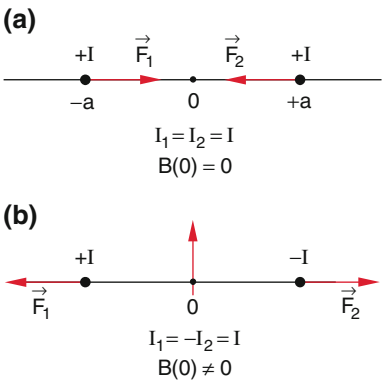


Fig. A.12 **a**) and **b**): Illustration of solution 3.1a

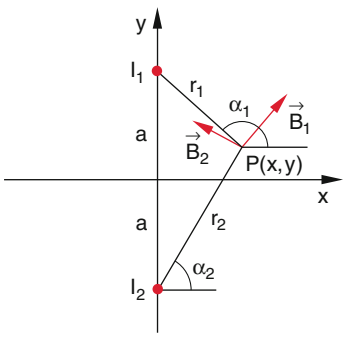


Fig. A.13 Illustration of solution 3.1b

with $r_1^2 = x^2 + (y-a)^2$; $r_2^2 = x^2 + (y+a)^2$.
Special cases:(α) $I_1 = I_2 = I$, $y = 0$ (field on the x-axis)

$$B_x = 0; \quad B_y = \frac{\mu_0 I}{\pi} \frac{x}{a^2 + x^2} = |\boldsymbol{B}|.$$

On the y-axis ($x = 0$) outside of the wires ($y \neq \pm a$) is

$$B_x = \frac{\mu_0 I}{\pi} \frac{y}{a^2 - y^2}; \quad B_y = 0$$
$$\Rightarrow |B| = B_x.$$

(β) $I_1 = -I_2 = I$: Now we obtain for $y = 0$:

$$B_x = \frac{\mu_0 I}{\pi} \frac{a}{a^2 + x^2}; \quad B_y = 0$$

and on the y-axis ($x = 0$) for $y \neq \pm a$

$$B_x = \frac{\mu_0 I}{\pi} \frac{a}{a^2 - y^2}; \quad B_y = 0.$$

(c) For parallel wires the force between the conductors per meter length is according to (3.32)

$$\frac{F}{L} = \frac{\mu_0}{4\pi a} I_1 \cdot I_2 (\hat{e}_\varphi \times \hat{e}_z),$$

where $\hat{e}_z$ is the unit vector pointing into the z-direction $\hat{e}_\varphi$ gives the azimuthal direction of the magnetic field B of a wire at the position of the other wire (Fig. A.11). For $I_1 = I_2 = I$ the forces F_1 and F_2 are antiparallel towards each other (Fig. A.12a) (Attraction), For $I_1 = -I_2 = I$ they are antiparallel away from each other (Fig. A.12b) (repulsion). The amount of the forces is for both cases

$$\frac{|\boldsymbol{F}|}{L} = \frac{\mu_0 I^2}{4\pi a}.$$

(d) In case of two perpendicular wires where one wire points into the x-direction the other into the z-direction with $y = -a = -2$ cm the magnetic field generated by the wire in x-direction is $B_1 = \{0, B_y, B_z\}$, the field induced by the wire in z-direction is $B_2 = \{B_x, B_y, 0\}$. The force exerted by the field B_1 on the length element dL of the wire in z-direction is

$$d\boldsymbol{F} = I_2(d\boldsymbol{L} \times \boldsymbol{B}_1)$$
$$d\boldsymbol{L} = \{0, 0, dz\}; \qquad \boldsymbol{B}_1 = \{0, B_y, B_z\}$$
$$\Rightarrow dF_x = -I_2 B_y dz; \quad dF_y = dF_z = 0.$$

The component B_y of the magnetic field caused by the wire in x-direction is at the point $P(0, -a, z)$ of the wire in z-direction is

$$B_y = \frac{\mu_0 I_1}{2\pi} \frac{z}{a^2 + z^2};$$
$$\Rightarrow dF_x = \frac{\mu_0}{2\pi} I_1 I_2 \frac{z dz}{a^2 + z^2}.$$

The force on the length element Δz from $z_1 = -b$ to $z_2 = +b$ is

$$F_x = \int_{z_1}^{z_2} \mathrm{d}F_x = \frac{\mu_0 I_1 I_2}{4\pi} \ln\left(a^2 + z^2\right)\Big|_{z=-b}^{z=+b} = 0.$$

The force between the two wires is therefore zero!
Question: Could this result be obtained simply by symmetry consideration?
Answer: Yes. Establish this answer.

3.2 Because of the cylinder symmetry the magnetic field has only a tangential component $B_\varphi(r)$, which can be calculated from

$$\int \boldsymbol{B} \cdot \mathrm{d}s = 2\pi r \cdot B_\varphi = \mu_0 \cdot I(r),$$

where $I(r)$ is the current through a cross section along the integration path. We then obtain:

(1) $r \leq r_1$: $\Rightarrow B = 0$;
(2) $r \geq r_4 \Rightarrow B = 0$ because the total current $I = I_1 + I_2$ with $I_2 = -I_1$ is zero.
(3) $r_1 \leq r \leq r_2$.

$$B = \frac{\mu_0 I}{2\pi r}\left(\frac{r^2 - r_1^2}{r_2^2 - r_1^2}\right);$$

(4) $r_2 \leq r \leq r_3$

$$B = \frac{\mu_0 I}{2\pi r};$$

(5) $r_3 \leq r \leq r_4$

$$B = \frac{\mu_0 I}{2\pi r}\left(1 - \frac{r^2 - r_3^2}{r_4^2 - r_3^2}\right).$$

3.3 The motion of the electron corresponds to the current

$$I = -e \cdot \nu = -e \cdot \omega/2\pi.$$

The frequency ω of the circular motion is obtained from

$$m\omega^2 \cdot r = \frac{1}{4\pi\varepsilon_0}\frac{e^2}{r^2},$$

because the centripetal force is equal to the Coulomb force.

$$\omega = \left(\frac{e^2}{4\pi\varepsilon_0 m r^3}\right)^{1/2}$$

$$\Rightarrow I = -\frac{e^2}{2\pi}\sqrt{\frac{1}{4\pi\varepsilon_0 m r^3}} \approx 1\,\mathrm{mA}.$$

The magnetic field at the center of the circular motion is according to (3.19a)

$$B_z = \frac{\mu_0 \cdot I}{2r} = -\frac{\mu_0 e^2}{4\pi r}\sqrt{\frac{1}{4\pi\varepsilon_0 m r^3}} \approx 12.5\,\mathrm{T}.$$

3.4 From (3.31) we get for the force onto the path element $\mathrm{d}L$ of the conductor with current I

$$\mathrm{d}\boldsymbol{F} = I \cdot (\mathrm{d}\boldsymbol{L} \times \boldsymbol{B})$$

$$\mathrm{d}\boldsymbol{L} = \begin{Bmatrix} \mathrm{d}x \\ \mathrm{d}y \\ 0 \end{Bmatrix} = \begin{Bmatrix} r \cdot \sin\varphi\,\mathrm{d}\varphi \\ r \cdot \cos\varphi\,\mathrm{d}\varphi \\ 0 \end{Bmatrix}$$

Since $\mathbf{B} = \{0, 0, B\}$ has only a z-component we get

$$\mathrm{d}\boldsymbol{F}_x = I \cdot \mathrm{d}y \cdot B$$

$$\mathrm{d}F_y = -I \cdot \mathrm{d}x \cdot B$$

$$\Rightarrow F_x = I \cdot B \cdot r \cdot \int_0^\pi \cos\varphi\,\mathrm{d}\varphi = 0$$

$$F_y = -I \cdot B \cdot r \cdot \int_0^\pi \sin\varphi\,\mathrm{d}\varphi = -2r \cdot I \cdot B.$$

The same force would act onto a straight wire of length $L = er$ which carries the current I.

3.5 (a) According to (3.22b) the magnetic field at $z = 0$ is

$$B(z=0) = \frac{\mu_0 N I R^2}{\left[(d/2)^2 + R^2\right]^{3/2}}.$$

With $N = 100$, $R = 0.4$ m we obtain

$$B(z=0) = \mu_0 I \frac{16\,\mathrm{m}^2}{\left[0.16\,\mathrm{m}^2 + (d/2)^2\right]^{3/2}}.$$

For $d = R$ and $I = 1$ A this becomes

$$B(z=0) = 2.25 \times 10^{-4}\,\mathrm{T} = 2.25\,\mathrm{Gauß}$$

(b) With $B(0) = 5 \times 10^{-5}$ T the current is $I = 0.22$ A. The coil axis has to be aligned antiparallel to the direction of the earth magnetic field.
(c) For the calculation of the field outside the solenoid we set $z = \pm(d/2 + \Delta z)$ where Δz is the external distance from the plane of the solenoid. Expanding (3.22a) into a Taylor series around $\Delta z = 0$ we get

$$B(z) = \frac{\mu_0 I R^2}{2}\left[\frac{1}{\left[(d+\Delta z)^2 + R^2\right]^{3/2}} + \frac{1}{(\Delta z^2 + R^2)^{3/2}}\right].$$

For $d = R$ this gives

$$B(z) = \frac{\mu_0 I}{2R}\left[\frac{1}{\left[1+\left(1+\frac{\Delta z}{R}\right)^2\right]^{3/2}} + \frac{1}{\left[1+\left(\frac{\Delta z}{R}\right)^2\right]^{3/2}}\right]$$
$$\approx \frac{\mu_0 I}{2R}\left[\frac{1}{\sqrt{8}}\left(1-\frac{3}{2}\frac{\Delta z}{R}-\frac{3}{4}\left(\frac{\Delta z}{R}\right)^2\right)\right.$$
$$\left.-\frac{15}{8}\left(\frac{\Delta z}{R}\right)^2+\cdots+1-\frac{3}{2}\frac{\Delta z}{R}-\frac{15}{8}\left(\frac{\Delta z}{R}\right)^2+\cdots\right]$$
$$\approx \frac{\mu_0 I}{2R}\left[1.35-2\frac{\Delta z}{R}-2.8\left(\frac{\Delta z}{R}\right)^2-\cdots\right].$$

3.6 (a) The determination of the electron trajectory in the magnetic field $\mathbf{B} = \{0, 0, B_0\}$ proceeds as follows:

The velocity component $v_z = v_0/\sqrt{3}$ remains constant. The other two components can be obtained from the condition: Lorentz-force is equal to the centripetal force:

$$e\cdot(\upsilon\times\boldsymbol{B}) = m\cdot\omega^2\cdot\begin{Bmatrix}x\\y\\0\end{Bmatrix}$$
$$\Rightarrow e\upsilon_y B_0 = m\omega^2 x,$$
$$-e\upsilon_y B_0 = m\omega^2 y.$$

With $r^2 = x^2 + y^2$ and $v_\perp^2 = v_x^2 + v_y^2$ if follows:

$$e^2\upsilon_\perp^2 B_0^2 = m^2\omega^4 r^2.$$

For $v_z = 0$ the electron follows a circular path in the x-y-plane with the radius

$$r = \frac{m\cdot\upsilon_\perp}{eB_0} = \frac{m\cdot\upsilon_0\cdot\sqrt{2}}{e\cdot B_0\cdot\sqrt{3}}.$$

The orbital period is

$$T = \frac{2\pi r}{\upsilon_\perp} = \frac{2\pi m}{eB_0}.$$

With $v_z = v_0\cdot\sqrt{3}$ the trajectory is a circular helix around the z-axis with a helix pitch

$$\Delta z = \upsilon_2\cdot T = \frac{2\pi\cdot\upsilon_0\cdot m}{\sqrt{3e}\cdot B_0}.$$

For this example the quantities v_z, $v_r = \dot{r} = 0$, $|v|$, $|p| = m\cdot|v|$ are temporally constant.

(b) An additional electric field $E_1 = E_0\{0, 0, 1\}$ affects only v_z but not v_x and v_y. It is

$$\upsilon_z = \upsilon_z(0) + a\cdot t = \upsilon_0/\sqrt{3} + \frac{eE_0}{m}t.$$

The electron trajectory remains s helix with a helix pitch, that increases with time. We obtain:

$$\Delta z(t) = \upsilon_z\cdot T = \left(\upsilon_0/\sqrt{3} + \frac{eE}{m}t\right)\frac{2\pi m}{eB_0}$$
$$= \Delta z_0 + \frac{2\pi E_0}{B_0}t.$$

Only $v_r = 0$ remains constant.
An additional electric field $E_2 = E_0\{1, 0, 0\}$ leads to the coupled differential equations

$$\ddot{x} = \frac{e}{m}E_0 + \frac{e}{m}B_0\dot{y},$$
$$\ddot{y} = -\frac{e}{m}B_0\dot{x},$$

With the initial conditions $\dot{x}(0) = \dot{y}(0) = v_0/\sqrt{3}$ the solutions are

$$\dot{x}(t) = \frac{\upsilon_0}{\sqrt{3}}\cos\omega t + \left(\frac{E_0}{B_0} + \frac{\upsilon_0}{\sqrt{3}}\right)\sin\omega t,$$
$$\dot{y}(t) = -\frac{E_0}{B_0} + \left(\frac{E_0}{B_0} + \frac{\upsilon_0}{\sqrt{3}}\right)\cos\omega t - \frac{\upsilon_0}{\sqrt{3}}\sin\omega t.$$

Integration of these equations gives the trajectories $x(t)$ and $y(t)$. None of the quantities given in (a) remains constant.

3.7 (a) The drift velocity of the electrons is given by the current density

$$j = n\cdot e\cdot\upsilon_D = I/A$$
$$\Rightarrow |\upsilon_D| = \frac{I}{n\cdot e\cdot A}$$
$$= \frac{10}{8\times10^{28}\cdot1.6\times10^{-19}\times10^{-4}\times10^{-2}}\frac{\mathrm{m}}{\mathrm{s}}$$
$$= 0.78\times10^{-3}\,\mathrm{m/s} = 0.78\,\mathrm{mm/s}.$$

(b) The Hall-voltage is according to (3.43c)

$$U_H = \frac{I\cdot B}{n\cdot e\cdot d}$$

with $d = \Delta y = 1$ cm, $B = 2$ T, $I = 10$ A, $n_e = 8\times10^{18}\,\mathrm{m}^{-3} \Longrightarrow U_H = 1.56\times10^{-7}$ V $= 0.156\,\mu$V.

(c) The force per meter of the copper rod is

$$\frac{F}{l} = I\cdot B = 10\cdot2\,\mathrm{N/m} = 20\,\mathrm{N/m}.$$

3.8 (a) The electric resistance of the iron yoke is

$$R_{\mathrm{Fe}} = \varrho \cdot \frac{L}{A} = 8.71 \times 10^{-8} \cdot \frac{0.6}{5 \times 10^{-6}}\,\Omega$$
$$= 1.05 \times 10^{-2}\,\Omega.$$
$$R_{\mathrm{Konst}} = \frac{0.5 \times 10^{-6} \cdot 0.2}{5 \times 10^{-6}\Omega}$$
$$= 2 \times 10^{-2}\Omega$$
$$U_{\mathrm{th}} = a \cdot \Delta T = 53 \times 10^{-6} \cdot (750 - 15)\mathrm{V}$$
$$= 39\,\mathrm{mV}.$$

The current through the circuit is then

$$I_{\mathrm{th}} = \frac{U_{\mathrm{th}}}{R_{\mathrm{Fe}} + R_{\mathrm{Konst}}} = \frac{3.9 \times 10^{-2}}{3.05 \times 10^{-2}}\mathrm{A}$$
$$= 1.28\,\mathrm{A}.$$

(b) The magnetic field at the center of the quadratic loop in the x-y-plane with side length a = 20 cm has only a z-component. Integrating (3.17) only from $-\pi/4$ to $+\pi/4$ gives the magnetic field generated by one side of the quadratic loop.

$$B_1 = \frac{\mu_0 I}{4\pi a/2} \int_{-\pi/4}^{\pi/4} \cos\alpha\,\mathrm{d}\alpha = \frac{\mu_0 I}{\sqrt{2}\pi a},$$

This gives for the total field

$$B = 4B_1 = \frac{2\sqrt{2}\mu_0 I}{\pi a} = 7.2 \times 10^{-6}\,\mathrm{T}.$$

If the current loop is embedded into a ferromagnetic material (e.g. Perm alloy with $\mu = 10^4$) one can reach B = 0.07 T (Fig. A.13).

3.9 For the Wien-filter (Fig. A.14) particles with the velocity v_0 can pass the filter if the condition

$$v_0 \cdot q \cdot B = q \cdot E \Rightarrow v_0 = \frac{E}{B}.$$

is fulfilled. Particles with the velocity $v = v_0 + \Delta v$ experience the additional force

$$\Delta F = \Delta v \cdot q \cdot B = m \cdot \ddot{x}$$
$$\Rightarrow \frac{\mathrm{d}x}{\mathrm{d}t} = \frac{q}{m}\Delta v \cdot B \cdot t + C_1.$$

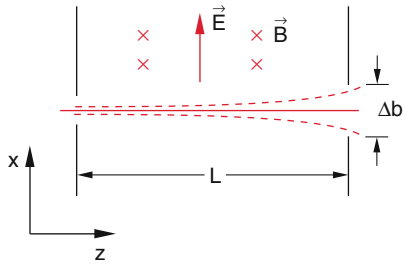


Fig. A.14 Illustration of solution 3.9

For particles which enter the field at t = 0 with the velocity in z-direction is $(\mathrm{d}x/\mathrm{d}t)_{t=0} = 0 \Rightarrow C_1 = 0$. Integration gives

$$x = \frac{1}{2}\frac{a}{m}\Delta v \cdot B \cdot t^2 + C_2.$$

If $x(t = 0) = 0 \Rightarrow C_2 = 0$. The transit time through the filter with length L is

$$t = \frac{L}{v} \approx \frac{L}{v_0} \Rightarrow \Delta v = \frac{2m \cdot x \cdot v_0^2}{q \cdot B \cdot L^2}.$$

For $x \leq \Delta b/2$ is

$$|\Delta v| \leq \frac{m \cdot \Delta b \cdot v_0^2}{q \cdot B \cdot L^2}.$$

Chapter 4

4.1 The conducting rod is dragged with constant velocity v over the yoke with width b (Fig. A.15).
The induced voltage is then

$$U_{\mathrm{ind}} = -\frac{\mathrm{d}\phi}{\mathrm{d}t}$$
$$= -B \cdot \frac{\mathrm{d}F}{\mathrm{d}t} = -B \cdot b \cdot v.$$

(a) The moving conducting rod represents the current

$$I = \varrho_{\mathrm{el}} \cdot b \cdot d \cdot v$$

(d = thickness of the conductive wire of the yoke). The current density is

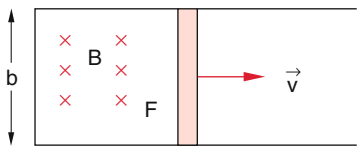


Fig. A.15 Illustration of solution 4.1

$$j = \varrho_{\mathrm{el}} \cdot v = -n \cdot e \cdot v$$

The induced voltage is then with $b \cdot v = -I/(n \cdot e \cdot d)$

$$U_{\mathrm{ind}} = \frac{I \cdot B}{n \cdot e \cdot d},$$

This is identical with the Hall voltage (3.43c).

(b) The mechanical power is

$$\frac{\mathrm{d}W_{\mathrm{mech}}}{\mathrm{d}t} = \text{Lorentz-force times velocity.}$$

The Lorentz force is according to (3.31) $F_{\mathrm{Lor}} = I \cdot b \cdot B$. We therefore get

(c) $$\frac{\mathrm{d}W_{\mathrm{mech}}}{\mathrm{d}t} = I \cdot b \cdot B \cdot v = -I \cdot U_{\mathrm{ind}}$$

$$\begin{aligned} U_{\mathrm{ind}} &= -\frac{\mathrm{d}}{\mathrm{d}t}\int \boldsymbol{B} \cdot \mathrm{d}\boldsymbol{F} \\ &= -\frac{\mathrm{d}}{\mathrm{d}t}\int a \cdot x \cdot b \cdot \mathrm{d}x \\ &= -a \cdot b \cdot \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{x^2}{2}\right) \\ &= -a \cdot b \cdot x \cdot v \\ x = v \cdot t &\Rightarrow U_{\mathrm{ind}} = ab \cdot v^2 \cdot t. \end{aligned}$$

The electric resistance of the yoke is

$$R(t) = (2b + 2x)g = 2g(b + v \cdot t).$$

The current is then

$$I(t) = \frac{U(t)}{R(t)} = \frac{a \cdot b \cdot v^2 \cdot t}{2g(b + v \cdot t)}.$$

4.2 We assume at first that the distance between the concentric tubes is large compared to the wall thickness of the tubes. We then get for the magnetic field

$$B = \frac{\mu_0 I}{2\pi r} \quad \text{for} \quad R_1 \le r \le R_2.$$

The magnetic flux ϕ through the rectangular cross section $A = a \cdot b$ with $a = R_2 - R_1$ and $b = l$ is

$$\phi = \frac{\mu_0 I \cdot l}{2\pi}\int_{R_1}^{R_2} B \cdot \mathrm{d}r = \frac{\mu_0 I \cdot l}{2\pi}\ln\frac{R_2}{R_1}.$$

(a) The inductance per m length of the tubes is

$$\widehat{L} = \frac{\mu_0}{2\pi}\ln\frac{R_2}{R_1}.$$

Numerical example: $R_1 = 1$ mm, $R_2 = 5$ mm

$$\Rightarrow \widehat{L} = \frac{1.26 \times 10^{-6}}{2\pi}\ln 5\,\mathrm{H/m} = 0.32 \times 10^{-6}\,\mathrm{H/m}.$$

(b) The energy density is

$$w(r) = \frac{1}{2}\frac{B^2}{\mu_0} = \frac{1}{2\mu_0}\frac{\mu_0^2 I^2}{4\pi^2 r^2} = \frac{\mu_0 I^2}{8\pi^2 r^2}.$$

The energy is then

$$\begin{aligned} W &= \int w\,\mathrm{d}v = 2\pi l \int_{R_1}^{R_2} w(r) r \mathrm{d}r \\ &= \frac{\mu_0 I^2 l}{4\pi}\ln\frac{R_2}{R_1} = \frac{1}{2}LI^2. \end{aligned}$$

The energy per m length is

$$\widehat{W} = \frac{1}{2}\widehat{L}I^2 = \frac{\mu_0 I^2}{4\pi}\ln\frac{R_2}{R_1}.$$

For a current of 10 A and $R_1 = 1$ mm, $R_2 = 5$ mm this gives

$$\widehat{W} = 1.6 \times 10^{-5}\,\mathrm{J/m}.$$

(c) If the wall thickness is not negligible the magnetic field in the inner tube must be calculated according to (3.9). This gives as additional contribution to the inductance per meter

$$L_2 = \frac{\mu\mu_0}{8\pi}$$

and for the energy per meter

$$\widehat{W} = \frac{\mu\mu_0 I^2}{16\pi}.$$

The calculation of the contribution of the outer tube leads to an integral that can be solved by a Taylor expansion of the integrand.

4.3 (a) The mutual inductance is according to (4.17)

$$L_{12} = \frac{\mu_0}{4\pi}\int_{s_1}\int_{s_2}\frac{\mathrm{d}s_1 \cdot \mathrm{d}s_2}{r_{12}},$$

With $\mathrm{d}s_1 \cdot \mathrm{d}s_2 = R_1 R_2 \mathrm{d}\varphi_1 \mathrm{d}\varphi_2 \cos(\varphi_1 - \varphi_2)$, and $r_{12} = \sqrt{R_1^2 + R_2^2 - 2R_1 R_2 \cos(\varphi_1 - \varphi_2)}$ we get

$$\Rightarrow L_{12} = \frac{\mu_0 \cdot R_1 R_2}{4\pi} \cdot \int_{\varphi_1=0}^{2\pi} \int_{\varphi_2=0}^{2\pi} \frac{\cos(\varphi_1 - \varphi_2)\mathrm{d}\varphi_1 \mathrm{d}\varphi_2}{\sqrt{R_1^2 + R_2^2 - 2R_1 R_2 \cos(\varphi_1 - \varphi_2)}}$$

$$= \frac{\mu_0}{4\pi} \frac{R_1 R_2}{\sqrt{R_1^2 + R_2^2}} \cdot \int_{\varphi_1=0}^{2\pi} \int_{\varphi_2=0}^{2\pi} \frac{\cos(\varphi_1 - \varphi_2)\mathrm{d}\varphi_1 \mathrm{d}\varphi_2}{\sqrt{1 - k \cdot \cos(\varphi_1 - \varphi_2)}}$$

with $k = 2R_1R_2/(R_1^2 + R_2^2)$.
With the substitution $\cos(1/2(\varphi_1 - \varphi_2)) = \sin\psi$ this leads to a sum of elliptical integrals which are listed e.g. in Bronstein's integral tables. For $R_1 \ll R_2 \Rightarrow k \ll 1$ the square root in the denominator can be expanded and the integral becomes

$$\int_{\varphi_1=0}^{2\pi} \int_{\varphi_2=0}^{2\pi} \cos(\varphi_1 - \varphi_2) \cdot \left[1 + \frac{1}{2} k \cos(\varphi_1 - \varphi_2)\right] \mathrm{d}\varphi_1 \, \mathrm{d}\varphi_2,$$

which has the solution $k\pi^2$: We then obtains for the inductance

$$L_{12} = \frac{\mu_0 \pi}{2} \frac{R_1^2 R_2^2}{\left[R_1^2 + R_2^2\right]^{3/2}}.$$

(b) The derivation in Sect. 4.3.2 and the Eq. (4.16)

$$\phi_m = \frac{\mu_0 I_1}{4\pi} \int_{s_1} \int_{s_2} \frac{\mathrm{d}s_1 \cdot \mathrm{d}s_2}{r_{12}}$$

is for $I_1 = I_2$ invariant under the interchange of the indices. The interchange of the indices $I_1 \rightarrow I_2$ and vice versa interchanges the situation that a current through the circuit 1 generates a magnetic field in the circuit 2 into the situation that a current through circuit 2 creates a magnetic field in the circuit 1. This means that $L_{12} = L_{21}$ (Fig. A.16).

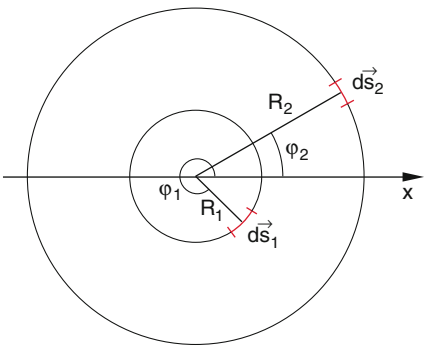


Fig. A.16 Illustration of solution 4.3

4.4 The capacity of the line with two metal stripes with width $2b$ and distance d between the stripes is

$$\widehat{C} = \varepsilon_0 \cdot \frac{2b}{d},$$

if vacuum is between the stripes. If material occupies the space between the stripes an additional factor ε has to be included. The calculation of the inductivity L is more tedious: We regard the magnetic field $\mathrm{d}\boldsymbol{B}$ at the point $P(x, y)$ produced by the current $\mathrm{d}I$ through an infinitesimal small band $\mathrm{d}x'$ of the metal stripe (Fig. A.17). With $\mathrm{d}I = I \cdot \mathrm{d}x'/(2b)$ we obtain

$$\mathrm{d}B = \frac{\mu_0 \mathrm{d}I}{2\pi r} = \frac{\mu_0 I}{4\pi \cdot b \cdot r} \mathrm{d}x'$$

with the components

$$\mathrm{d}B_x = -\frac{y}{r}\mathrm{d}B = -\frac{\mu_0 I}{2b\pi} \frac{y \cdot \mathrm{d}x'}{(x - x')^2 + y^2},$$

$$\mathrm{d}B_y = -\frac{x - x'}{r}\mathrm{d}B = \frac{\mu_0 I}{4\pi b} \frac{(x - x')\mathrm{d}x'}{(x - x')^2 + y^2}.$$

The field B which is produced by the total current I through the total metal stripe is

$$B = \int_{x'=-b}^{x'=+b} \mathrm{d}B.$$

With the substitution $u = (x' - x)/y$ we get

$$B_x = \frac{\mu_0 I}{4\pi b} \int_{u_1}^{u_2} \frac{\mathrm{d}u}{1 + u^2}$$

$$= -\frac{\mu_0 I}{4\pi b} \left[\arctan\frac{b - x}{y} + \arctan\frac{b + x}{y}\right],$$

$$B_y = -\frac{\mu_0 I}{4\pi b} \int_{u_1}^{u_2} \frac{u\,\mathrm{d}u}{1 + u^2}$$

$$= -\frac{\mu_0 I}{8\pi b} \ln\frac{y^2 + (b + x)^2}{y^2 + (b - x)^2}.$$

For $b \gg y$ is

$$\arctan\frac{b - x}{y} \rightarrow \frac{\pi}{2} \cdot \mathrm{sig}\, y$$

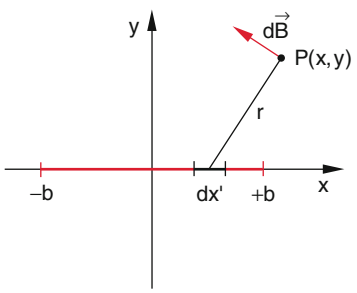


Fig. A.17 Illustration of solution 4.4

and

$$\ln\frac{y^2+(b+x)^2}{y^2+(b-x)^2}\to 4x/b.$$

This gives

$$B_x=-\frac{\mu_0 I}{4b}\operatorname{sig} y;\quad B_y=-\frac{\mu_0 I\cdot x}{2\pi\cdot b^2}.$$

For our double line is $y=\pm d/2$. This gives

$$B_x=-\frac{\mu_0 I}{4b}\cdot\operatorname{sig}(y\pm d/2).$$

The current through the upper stripe line is $+I$, through the lower line $-I$. The magnetic fields of the two lines point between the lines into the same direction namely the $+x$ direction. The two contributions therefore add. Outside the stripes the magnetic fields cancel each other. The magnetic field energy per m stripe length is

$$\widehat{W}_{\text{mag}}=\frac{1}{2\mu_0}B^2\cdot 2b\cdot d=\frac{B^2\cdot b\cdot d}{\mu_0}=\frac{\mu_0 I^2}{4b}\cdot d$$

with $B^2=B_x^2+B_y^2$.
Since $W_{\text{magn}}=1/2LI^2$ we get for the inductivity

$$\widehat{L}=\frac{\mu_0\cdot d}{2b}.$$

The product $C\cdot L$ of capacity C and inductivity L

$$\widehat{C}\cdot\widehat{L}=\varepsilon_0\mu_0$$

is independent of the geometrical dimensions of the double line as long as $d\ll b$.

4.5 The voltage induced in the pendulum is

$$U_{\text{ind}}=-\dot{\Phi}=-B\cdot \mathrm{d}F^*/\mathrm{d}t,$$

where $\mathrm{d}F^*/\mathrm{d}t$ is the area of the pendulum that enters the magnetic field per second.

$$\mathrm{d}F^*/\mathrm{d}t\propto v=L\cdot\dot{\varphi},$$

where L is the length of the pendulum from the pivot point to the center of the magnetic field.

$$\Rightarrow U_{\text{ind}}\propto\dot{\varphi}.$$

(a) The induced voltage generates eddy currents

$$I_{\text{e}}=U_{\text{ind}}/R,\Rightarrow I_{\text{e}}\propto \mathrm{d}\varphi/\mathrm{d}t$$

where R is the resistance for the eddy currents.

The damping torque $D_{\text{D}}=L\cdot F_{\text{L}}$ is determined by the Lorentz-force

$$|F_{\text{L}}|\propto I_{\text{W}}\cdot B$$

The force is according to Lenz's rule pointing in such a direction that it hinders the motion that generates the force. This means $D_D\propto-\dot{\varphi}$.

(b) Since $I_{\text{w}}\propto U_{\text{ind}}\propto B$ we get

$$D_D\propto B^2\propto I_{\text{F}}^2$$

where I_{F} is the field generating current.

4.6 The current I is

$$\begin{aligned}I(t)&=\frac{U_0}{R}(1-\mathrm{e}^{-(R/L)t})\\&=\frac{20}{100}(1-\mathrm{e}^{-(500t/\text{s})})\text{A}\\&=0.2(1-\mathrm{e}^{-(500t/\text{s})})\text{A}.\end{aligned}$$

At the time $t_0=0$ is $I(0)=0$, at the time $t_1=2$ ms is

$$\begin{aligned}I(t_1)&=0.2\left(1-\frac{1}{e}\right)A=0.126\,\text{A},\\I(\infty)&=0.2\,\text{A}.\end{aligned}$$

4.7 Gauss's law is for a vector function $\boldsymbol{u}(x,y,z)$

$$\oint\boldsymbol{u}\,\mathrm{d}S=\int\operatorname{div}\boldsymbol{u}\,\mathrm{d}V,$$

where S is the surface of the volume V. The conservation of the electric charge $Q=\int\varrho_{\text{el}}\mathrm{d}V$ demands

$$\begin{aligned}-\frac{\mathrm{d}Q}{\mathrm{d}t}&=-\frac{\mathrm{d}}{\mathrm{d}t}\int\varrho_{\text{el}}\mathrm{d}V=-\frac{\partial}{\partial t}\int\varrho_{\text{el}}\mathrm{d}V\\&=-\int\frac{\partial\varrho_{\text{el}}}{\partial t}\mathrm{d}V=\oint_S\varrho_{\text{el}}\upsilon\cdot\mathrm{d}\boldsymbol{S},\end{aligned}$$

since spatial and temporal integration can be interchanged. The partial differentiation $\partial\varrho/\partial t$ takes into account that $\varrho(x,y,z)$ can depend on the coordinates (x,y,z). However, the total charge Q inside the volume V does not depend on the coordinates, even if ϱ does depend on (x,y,z). Therefore the total derivative $\mathrm{d}Q/\mathrm{d}t$ is equal to the partial derivative $\partial Q/\partial t$. From

$$\int_S\varrho_{\text{el}}\upsilon\cdot\mathrm{d}S=\int\operatorname{div}(\varrho_{\text{el}}\upsilon)\mathrm{d}V$$

(Gauß's law) it follows the continuity equation

$$\operatorname{div} \boldsymbol{j} + \frac{\partial \varrho}{\partial t} = 0$$

with $\boldsymbol{j} = \varrho_{\mathrm{el}} \cdot \upsilon$

4.8 The train acts as short circuit bar. We therefore are faced with a problem that is equivalent to that of Problem 4.1

$$U_{\mathrm{ind}} = -B_{\perp} \cdot b \cdot \upsilon = -|B| \cdot \cos 65^{\circ} \cdot b \cdot \upsilon.$$

With b = 1.5 m, v = (200/3.6) m/s we get

$$U_{\mathrm{ind}} = 4 \times 10^{-5} \cdot \cos 65^{\circ} \cdot 1.5 \cdot \frac{200}{3.6}$$
$$= 1.41 \times 10^{-3}\,\mathrm{V} = 1.41\,\mathrm{mV}.$$

4.9 (a) If the straight wire is concentric to the circular loop with N windings (Fig. A.18) the magnetic field induced by the current through the straight wire is always directed along the circular loop. The magnetic flux $\mathrm{d}\Phi = B \cdot \mathrm{d}F$ is therefore zero because $\boldsymbol{B}$ is perpendicular to the vector $\mathrm{d}\boldsymbol{F}$ which points in the direction of the straight wire perpendicular to the circular loop area.

(b) the situation is different for the arrangement of Fig. A.18b.

The magnetic field of the straight wire is

$$B = \frac{\mu_0 I}{2\pi r}$$

and the magnetic flux through the rectangular coil cross section $F = a \cdot b$ is

$$\Phi = \int_F B\,\mathrm{d}F = \frac{b \cdot \mu_0 I}{2\pi} \int_{r=d}^{d+a} \frac{\mathrm{d}r}{r}$$
$$= \frac{\mu_0 \cdot b \cdot I}{2\pi} \ln \frac{d+a}{d} = \frac{\mu_0 \cdot b \cdot I}{2\pi} \ln\left(1 + \frac{a}{d}\right).$$

With $I = I_0 \cdot \sin \omega t$ is

$$U_{\mathrm{ind}} - N \cdot \dot{\Phi} = U_0 \cdot \cos \omega t$$

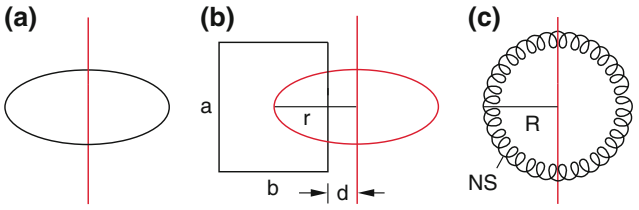


Fig. A.18 Illustration of solution 4.9

with

$$U_0 = \frac{N \cdot \omega \cdot I_0 \cdot \mu_0 \cdot b}{2\pi} \ln\left(1 + \frac{a}{d}\right).$$

(c) For the toroid coil in Fig. A.18c the toroid windings enclose the magnetic field lines. For a radius r_{s} of the coil windings the coil cross section is $F = N \cdot \pi \cdot r_{\mathrm{S}}^2$. The magnetic flux is then with $\xi = \sqrt{r_{\mathrm{S}}^2 - z^2}$)

$$\Phi = \int B\,\mathrm{d}F$$
$$= \frac{N \cdot \mu_0 I}{2\pi} \int_{z=-r_{\mathrm{S}}}^{+r_{\mathrm{S}}} \left(\int_{r=R-\xi}^{R+\xi} \frac{\mathrm{d}r}{r} \right) \mathrm{d}z$$
$$= \frac{N \cdot \mu_0 I}{2\pi} \int_{z=-r_{\mathrm{S}}}^{+r_{\mathrm{S}}} \ln \frac{R+\xi}{R-\xi}\,\mathrm{d}z$$
$$= \frac{N \cdot \mu_0 I}{2\pi} \int_{z=-r_{\mathrm{S}}}^{+r_{\mathrm{S}}} \left[\ln\left(\mathrm{R} + \sqrt{r_{\mathrm{S}}^2 + z^2}\right) - \ln\left(\mathrm{R} - \sqrt{r_{\mathrm{S}}^2 + z^2}\right) \right] \mathrm{d}z.$$

4.10 The magnetic field in the iron kernel is

$$B = \mu \cdot \mu_0 \cdot n \cdot I = 1\,\mathrm{T}$$

With $n = N/l$ is

$$\Rightarrow \mu = \frac{B}{\mu_0 \cdot \mu \cdot I} = \frac{0.4}{4\pi \cdot 10^{-7} \cdot 10^{-3}} = 320.$$

The inductivity is then

$$L = \mu \cdot \mu_0 \cdot n^2 F \cdot l = 10\,\mathrm{H}.$$

If the external circuit is switched off within 1 ms, the induced voltage is

$$U_{\mathrm{ind}} = -L \cdot \frac{\mathrm{d}I}{\mathrm{d}t} = -10 \times 10^{-3}\mathrm{V} = -10\,\mathrm{kV}.$$

The output voltage current jumps from the value $I(t < 0) = U/R$ to the value

$$I(t > 0) = I_0 = \frac{U_{\mathrm{ind}}}{R_2} = \frac{10 \cdot 10^3}{5}\,\mathrm{A} = 2000\,\mathrm{A}.$$

it decreases the exponentially as

$$I = I_0 \cdot \mathrm{e}^{-(R/2)t}$$

The situation is similar to that in Fig. 4.14c-d.

Chapter 5

5.1 (a) R and C must be connected in parallel (Fig. A.19).

$$Z_1 = R, Z_2 = \frac{1}{\mathrm{i}\omega C}$$
$$\Rightarrow Z = \frac{Z_1 \cdot Z_2}{Z_1 + Z_2} = \frac{R}{\mathrm{i}\omega C\left(R + \frac{1}{\mathrm{i}\omega C}\right)}$$
$$= \frac{R}{1 + \mathrm{i}\omega RC}$$
$$|Z| = \frac{R}{\sqrt{1 + (\omega RC)^2}}$$

$$|Z(\omega = 0)| = R = 100\,\Omega$$
$$|Z(\omega = 2\pi \cdot 50/s)| = 20\,\Omega$$
$$= \frac{100}{\sqrt{1 + 4\pi^2 \cdot 2500 \cdot 100^2 \cdot C^2}}$$
$$\Rightarrow C = 156\,\mu\mathrm{F}.$$

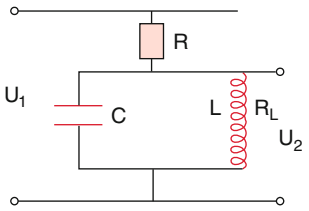


Fig. A.19 Illustration of solution 5.1

(b) Since the output voltage U for $\omega = 0$ is $U \neq 0$ the circuit must be a parallel circuit. For $\omega = 0$ is

$$\frac{U_2}{U_1} = \frac{R_\mathrm{L}}{R + R_\mathrm{L}} = 0.01$$
$$\Rightarrow R = \frac{0.99 R_\mathrm{L}}{0.01} = 99 R_\mathrm{L} = 99\,\Omega.$$

The maximum output voltage is obtained for $\omega \cdot L - 1/(\omega C) = 0$, i.e. for the resonant frequency

$$\omega_\mathrm{R} = \frac{1}{\sqrt{LC}} \Rightarrow C = \frac{1}{(L\omega_\mathrm{R}^2)} = 1.78\,\mathrm{mF}.$$

The approximation $\omega_\mathrm{R} = 1\sqrt{L \cdot C}$ is only valid for small values of the resistance R_L.
If R_L cannot be neglected one has to form the derivative of

$$\frac{|U_2|}{|U_1|} = \left| 1 - \frac{R}{R + \frac{1}{i\omega C + \frac{1}{1\omega L + R_L}}} \right|$$

with respect to ω and then set this derivative = 0 (extreme value of the function). The corresponding equation must be solved for C. For $R_\mathrm{L} = 1\ \Omega$ this gives $C = 1.80$ mF, for $R_\mathrm{L} = 20\ \Omega$ one obtains $C = 5.15$ mF.

Remark Although such calculations train the brain, they are more suited for computer calculations rather than for physicists who want to illustrate the physical essence of problems.

5.2 The resistance of the complete circuit in Fig. 5.35a is the sum

$$Z_\mathrm{tot} = Z_\mathrm{K} + R,$$

where

$$Z_K = \frac{Z_1 \cdot Z_2}{Z_1 + Z_2}$$

with

$$Z_1 = \frac{1}{\mathrm{i}\omega C}; \quad Z_2 = \mathrm{i}\omega L + R_\mathrm{L}$$

here Z_tot is the resistance of the parallel circuit and R is the load resistor which is here assumed as pure Ohmic resistor. The output voltage is then

$$U_\mathrm{a} = \frac{R}{Z_\mathrm{K} + R} U_\mathrm{e} = \frac{R}{Z_\mathrm{tot}} \cdot U_\mathrm{e}.$$

For the complex resistance Z_tot we obtain

$$Z_\mathrm{K} = \frac{R_\mathrm{L} + \mathrm{i}\omega L}{(1 - \omega^2 LC) + \mathrm{i}\omega R_\mathrm{L} C},$$

So, look for the total resistance

$$Z_\mathrm{tot} = \frac{R_\mathrm{L} + R - \omega^2 RLC + \mathrm{i}\omega(L + R_\mathrm{L} RC)}{(1 - \omega^2 LC) + \mathrm{i}\omega R_\mathrm{L} C}$$

The resonance frequency of the undamped parallel circuit is with $L = 10^{-4}$ H and $C = 10^{-6}$ F

$$\omega_\mathrm{R} = \frac{1}{\sqrt{L \cdot C}} = 10^5\ \mathrm{s}^{-1}.$$

Since the inductive resistance $|\omega_\mathrm{R} \cdot L| = 10\ \Omega$ at the resonance frequency ω_R is large compared with the Ohmic resistance $R_\mathrm{L} = 1\ \Omega$ of the inductance the resonance frequency of the damped circuit is only by 1% smaller than that of the undamped circuit. The total resistance at the resonant frequency is

$$Z_\mathrm{tot}(\omega_\mathrm{R}) = R + \frac{L}{C \cdot R_\mathrm{L}} - \mathrm{i}\sqrt{L/C}.$$

Numerical values: $R_\mathrm{L} = 1\,\Omega, R = 50\,\Omega, C = 1\,\mu\mathrm{F}, L = 10^{-4}\,\mathrm{H}$

$$\Rightarrow Z_\mathrm{tot} = (150 - 10\mathrm{i})\Omega$$

with the amount

$$Z_{\text{tot}} = 150.3\,\Omega.$$

Note, that the total resistance at the resonance frequency is not real. This means that the output voltage U_a has a phase shift against the input voltage U_e. It is

$$\begin{aligned} U_A &= U_e(R/Z_{\text{tot}}) = U_e \\ &= U_e \cdot (0.332 + 0.022\text{i}) \\ \Rightarrow U_a &= U_e \cdot \cos(\omega t + \varphi). \end{aligned}$$

$\Longrightarrow$ With $\tan\varphi = 10/150 = 0.067 \Longrightarrow \varphi = 3.81°$.
The frequency dependence of the resistance Z_K of the parallel circuit can be determined by setting $R = 0$.
The frequency width of the resonance is approximately

$$\Delta\omega = \frac{R}{L} = 10^4\,\text{s}^{-1}.$$

This can be also expressed by the quality factor

$$Q = \frac{\omega L}{R} = 10$$

of the circuit, because the relation

$$\frac{\Delta\omega}{\omega_0} = \frac{1}{Q} = \frac{1}{10} \Rightarrow \Delta\omega = \frac{\omega_0}{10} = 10^4\,\text{s}^{-1}.$$

The frequencies ω_1 and ω_2 where the resistance Z has dropped to $½Z(\omega_R)$ are

$$\omega_{1.2} = \left(10^5 \pm 10^4\right)\text{s}^{-1}.$$

The full half width of the resonance curve is therefore $\Delta\omega = 2 \times 10^4\ \text{s}^{-1}$.

5.3 Since the total magnetic flux penetrates also the secondary coil the coupling factor is $k = 1$. Therefore the phase shift between U_2 and U_1 is $\Delta\varphi = 180°$ if both coils have the same winding orientation.

$$\Rightarrow \frac{U_2}{U_1} = -\frac{N_2}{N_1}.$$

(a) If the load resistor R is pure Ohmic, the ratio U_2/U_1 is independent of R as long as R is large compared with the resistance of the secondary coil. The effective input power is

$$\overline{P}_e = \frac{U_2^2}{R} = \left(\frac{N_2}{N_1}\right)^2 \frac{U_1^2}{R}.$$

The secondary current is according to (5.50b) with $L_{12} = \sqrt{(L_1 \cdot L_2)}$

$$I_2 = \frac{U_1}{R}\sqrt{\frac{L_2}{L_1}} = \frac{U_1}{R} \cdot \frac{N_2}{N_1} \Rightarrow \overline{P}_e = U_2 \cdot I_2.$$

(b) For a capacitive load and a coupling factor $k = 1$ is

$$\begin{aligned} \frac{U_2}{U_1} &= \frac{L_{12}}{L_1 - \omega^2 C_1 L_2 (1 - k^2)} \\ &= \frac{\sqrt{L_1 \cdot L_2}}{L_1} = \sqrt{\frac{L_2}{L_1}} = N_2/N_1 \end{aligned}$$

in this case the same result is obtained as for a pure Ohmic load.

5.4 From Fig. A.20, which represents a redrawing of Fig. 5.54 the following relations can be obtained:

$$\begin{aligned} Z_D &= \frac{1}{\text{i}\omega C} + \frac{1}{\frac{1}{\text{i}\omega L} + \frac{1}{R}} \\ Z_B &= \frac{1}{\text{i}\omega C} + \frac{1}{\frac{1}{\text{i}\omega L} + \frac{1}{Z_D}} \\ &= \frac{1}{\text{i}\omega C} + \frac{1}{\frac{1}{\text{i}\omega C} + \frac{1}{\text{i}\omega C} + \frac{1}{1/(\text{i}\omega L) + 1/R}} \\ Z &= \frac{1}{\frac{1}{1\omega L} + \frac{1}{Z_B}} \\ &= \frac{1}{\frac{1}{\text{i}\omega L} + \frac{1}{\frac{1}{\text{i}\omega L} + \frac{1}{\frac{1}{1\omega C} + \frac{1}{1/(1\omega L) + 1/R}}}} \end{aligned}$$

$U_a = U_1$, $I_a = U_1/(i \cdot \omega \cdot L)$, $I_B = I_1 - I_a$, $U_B = I_B \cdot Z_B$, $I_C = U_B/(i \cdot \omega \cdot L)$, $I_D = I_B - I_C$, $U_D = I_D \cdot Z_D = U_2$, $I_2 = U_D/R$, $I_1 = U_1/Z$.

Inserting these relations into the equation for Z gives

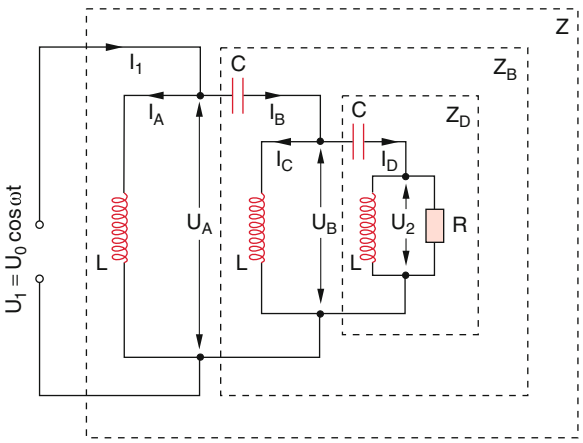


Fig. A.20 Illustration of solution 5.4

$$Z = (37.6 + 38.9\mathrm{i})\Omega, \quad |Z| = 54.1\Omega,$$
$$Z_B = (22.7 - 35.4\mathrm{i})\Omega, \quad |Z_B| = 42.0\Omega,$$
$$Z_D = (13.2 - 11.3\mathrm{i})\Omega, \quad |Z_D| = 17.4\,\Omega,$$
$$\frac{|U_2|}{|U_1|} = 0.414, \quad \frac{|I_2|}{|I_1|} = 0.448.$$

5.5 $\overline{P}_{el} = \overline{I \cdot U} = \overline{U_{ind}^2}/(R_i + R_a)$, because $I = U_{ind}/(R_i + R_a)$.

$$U_{ind} = -\frac{d\Phi}{dt} \cdot N = -B \cdot N \cdot F \cdot \omega \cdot \cos \omega t$$
$$\Rightarrow \overline{P}_{el} = \frac{1}{2}\frac{B^2 N^2 F^2 \omega^2}{R_i + R_a}$$
$$= \frac{1}{2}\frac{0.2^2 \cdot 25 \times 10^4 \times 10^{-4} \cdot 4\pi^2 \cdot 50^2}{10 + 5}\,\mathrm{kW}$$
$$= 3.29\,\mathrm{kW}.$$

5.6 The time constant of the capacitor discharge is

$$\tau = R \cdot C = 50 \times 10^{-3}\,\mathrm{s} = 50\,\mathrm{ms}.$$

$\Longrightarrow$ The discharge starts at $t = 0$ after the peak voltage U_0 has been reached.

$\Longrightarrow$ (a) One way rectification: The discharge lasts until the intersection of the curve $U_1(t) = U_0 \cdot e^{-t/(RC)}$ with the curve $U_2(t) = U_0 \cdot \cos(\omega t - 2\pi)$ (Fig. A.21a). This gives

$$t = -RC \cdot \mathrm{In}(\cos \omega t - 2\pi)$$
$$\Rightarrow t_1 = 17.5\mathrm{ms}, U(t_1 = 17.5\mathrm{ms})$$
$$= U_0 \cdot e^{-17.5/50} \approx 0.7 U_0.$$

The ripple of the DC output voltage is then

$$w = \frac{U_{max} - U_{min}}{U_{max}} = 0.3.$$

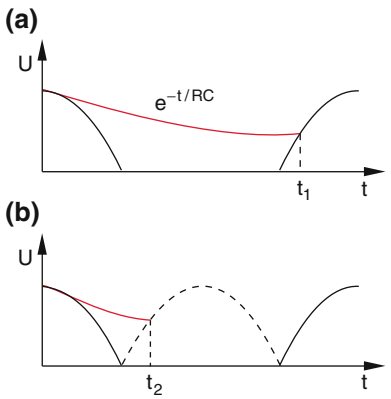


Fig. A.21 Illustration of solution 5.6

(b) For the Graetz-rectifier (full wave bridge circuit) one obtains (Fig. A.21b):

$$e^{-t_2/(RC)} = |\cos(\omega t - \pi)|$$
$$\Rightarrow t_2 = 8.3\,\mathrm{ms}$$
$$U = U_0 e^{-8.3/50} \Rightarrow \frac{U}{U_0} = 0.83$$
$$\Rightarrow w = 0.17.$$

With the Graetz- rectifier the ripple is smaller than that of the one-way rectification by the factor 0.17/0.3 = 0.57. Its frequency is, however, two times higher. It can be therefore much easier filtered by an RC-circuit (see Fig. 5.44).

$$Z = \frac{Z_1 \cdot Z_2}{Z_1 + Z_2} = \frac{R}{1 + \mathrm{i}\omega RC}.$$
$$I = \frac{U}{Z} = \frac{U_0 \cos \omega t}{R}(1 + \mathrm{i}\omega RC)$$
$$= \frac{U_0}{R}\sqrt{1 + \omega^2 R^2 C^2}\cos(\omega t + \varphi)$$
$$= I_0 \cos(\omega t + \varphi)$$

with

$$I_0 = \frac{U_0}{R}\sqrt{1 + \omega^2 R^2 C^2}$$

and

$$\tan \varphi = \frac{\omega RC}{1} = 2\pi \cdot 50 \times 10^7 \times 10^{-5}$$
$$= 3140 \Rightarrow \varphi \lesssim 90^\circ.$$

We then obtain

$$\overline{P}_{Wirk} = \overline{I \cdot U} = \frac{1}{2} I_0 U_0 \cos \varphi$$
$$\cos \varphi; = \frac{1}{\sqrt{1 + \tan^2 \varphi}} = \frac{1}{\sqrt{1 + (\omega CR)^2}}$$
$$\Rightarrow \overline{P}_{Wirk} = \frac{1}{2}\frac{U_0^2}{R}.$$

Only this part of the total power can be consumed. The rest is the wattless power

$$\overline{P}_{Blind} = \frac{1}{2} I_0 U_0 \sin \varphi = \frac{1}{2} U_0^2 \omega C.$$

Numerical values:

$$I_0 = 0.94\,\mathrm{A},$$
$$I_{Wirk_0} = 3 \times 10^{-5}\,\mathrm{A}$$
$$I_{Blind_0} = 0.94\,\mathrm{A}$$
$$\overline{P}_{Wirk} = 4.5\,\mathrm{mW}$$
$$\overline{P}_{Blind} = 141\,\mathrm{W}.$$

Although the wattless power does not produce Joule's heat it has to be taken into account for the dimensioning of the cable diameters (Fig. A.22).

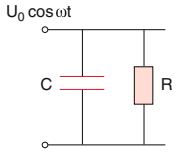


Fig. A.22 Illustration of solution 5.7

5.8 The current through the series circuit is

$$I = \frac{U_0 \sin \omega t}{Z} \quad \text{with} \quad \mathrm{Z} = \mathrm{R} + \mathrm{i}\left(\omega \mathrm{L} - \frac{1}{\omega \mathrm{C}}\right).$$

The voltage across the inductance is

$$\begin{aligned} U_\mathrm{L} &= \frac{\mathrm{i}\omega L}{Z} U_0 \sin \omega t \\ &= \frac{-\omega^2 LC}{1 - \omega^2 LC + \mathrm{i}\omega RC} U_0 \cdot \sin \omega t \\ &= -\frac{\omega^2 LC(1 - \omega^2 LC\mathrm{i}\omega RC)}{(1 - \omega^2 LC)^2 + \omega^2 R^2 C^2} U_0 \sin \omega t \\ &= U \cdot \sin(\omega t - \varphi) \end{aligned}$$

with

$$U = \frac{\omega^2 LC}{\sqrt{(1 - \omega^2 LC)^2 + \omega^2 R^2 C^2}}$$

and

$$\tan \varphi = \frac{\omega RC}{1 - \omega^2 LC} = 0.417 \Rightarrow \varphi = 22.6^\circ.$$

With the numerical values given, the voltage is

$$U = 0.302\,\mathrm{V}.$$

5.9 The ratio of output voltage to input voltage is (Fig. A.23):

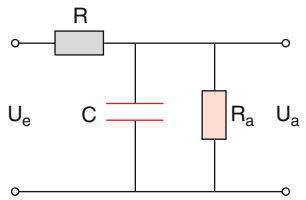


Fig. A.23 Illustration of solution 5.9

$$\frac{U_\mathrm{a}}{U_\mathrm{e}} = \frac{Z}{R + Z}.$$

where K is a real number.

$$\begin{aligned} Z &= \frac{R_\mathrm{a} \cdot \frac{1}{1\omega C}}{R_\mathrm{a} + \frac{1}{\mathrm{i}\omega C}} = \frac{R_\mathrm{a}}{1 + \mathrm{i}\omega R_\mathrm{a} C} \\ \frac{U_\mathrm{a}}{U_\mathrm{e}} &= \frac{R_\mathrm{a}}{R_\mathrm{a} + R + \mathrm{i}RR_\mathrm{a}\omega C} \\ &= \frac{R_\mathrm{a} \cdot (R_\mathrm{a} + R - \mathrm{i}RR_\mathrm{a}\omega C)}{(R_\mathrm{a} + R)^2 + (RR_\mathrm{a}\omega C)^2} \\ |U_\mathrm{e}| &= \frac{R_\mathrm{a}}{\sqrt{(R_\mathrm{a} + R)^2 + (RR_\mathrm{a}\omega C)^2}} \\ U_\mathrm{a} &= K \cdot U_\mathrm{e} \cdot \mathrm{e}^{\mathrm{i}\varphi}; \quad \tan \varphi = -\frac{RR_\mathrm{a}\omega C)}{R + R_\mathrm{a}} \end{aligned}$$

Numerical example: $R_\mathrm{a} = R = 1\,\mathrm{k\Omega}$, $C = 100\,\mu\mathrm{F}$,

(a) for $\omega = 0$:

$$\frac{|U_\mathrm{a}|}{|U_\mathrm{e}|} = \frac{R_\mathrm{a}}{R_\mathrm{a} + R} = \frac{1}{2};$$

(b) For $\omega = 2\pi \cdot 50\,\mathrm{Hz}$

$$\frac{|U_\mathrm{a}|}{|U_\mathrm{e}|} = 0.032.$$

5.10 The terminal voltage U_K is

$$U_\mathrm{K} = U_\mathrm{ind} - R_\mathrm{R}(I_\mathrm{F} + I_\mathrm{a}).$$

On the other hand is

$$U_\mathrm{K} = R_\mathrm{F} \cdot I_\mathrm{F}.$$

Comparing the two results gives for $U_\mathrm{ind} = U'_\mathrm{ind}$ and $I_\mathrm{F} = I_\mathrm{F2}$

$$U'_\mathrm{ind} = R_\mathrm{R} I_\mathrm{a} + (R_\mathrm{R} + R_\mathrm{F}) I_{\mathrm{F}_2}.$$

According to (5.6) is

$$U_\mathrm{K} = U'_\mathrm{ind} - R_\mathrm{R}(I_\mathrm{F} + I_\mathrm{a}).$$

Since U'_ind decreases with increasing load current (I_F2 becomes smaller) decreases also $U_\mathrm{K}(I_\mathrm{a})$ with increasing I_a. Therefore U_K becomes maximum for $I_\mathrm{a} = 0$.

Chapter 6

6.1 For the frequency ω the relation holds:

$$\omega = \sqrt{\frac{1}{LC} - \alpha^2} \quad \text{with} \quad \alpha = \frac{R}{2L},$$
$$\omega = 2\pi \cdot 8 \times 10^5\,\mathrm{s}^{-1} = 5 \times 10^6\,\mathrm{s}^{-1},$$
$$U = U_0 \cdot \mathrm{e}^{-\alpha t} \Rightarrow \alpha = \frac{1}{t}\ln\frac{U_0}{U}.$$

The oscillation period is then

$$T = \frac{2\pi}{\omega} = 1.25 \times 10^{-6}\mathrm{s}^{-1}.$$

After the time t = 30 T is $U/U_0 = 1/2$.

$$\Rightarrow \alpha = \frac{10^6}{30 \cdot 1.25}\ln 2 = 1.8 \times 10^4\,\mathrm{s}^{-1},$$
$$L = \frac{1}{C \cdot (\omega^2 + \alpha^2)} = \frac{10^9}{25 \times 10^{12} + 3.4 \times 10^8}\,\mathrm{H} \approx 4 \times 10^{-5}\,\mathrm{H},$$
$$\Rightarrow R = 2\alpha \cdot L = 2 \cdot 1.8 \times 10^4 \times 10^{-5} = 1.44\,\Omega.$$

6.2 The amount of the complex resistance of a series resonant circuit is

$$|Z| = \sqrt{R^2 + \left(\omega L - \frac{1}{\omega C}\right)^2} = \sqrt{R^2 + X^2}.$$

The ratio is then

$$\left|\frac{Z(\omega_0 + R/L)}{Z(\omega_0)}\right| = \frac{\sqrt{R^2 + X^2}}{\sqrt{R^2}} = \sqrt{1 + \frac{X^2}{R^2}},$$
$$X = \left(\omega_0 + \frac{R}{L}\right)L - \frac{1}{(\omega_0 + R/L)C}$$

With the resonant frequency $\omega_0 = 1/\sqrt{LC} \Rightarrow$

$$X = \sqrt{L/C} + R - \frac{1}{\sqrt{C/L} + RC/L} = R \cdot \left(1 + \frac{1}{1 + R \cdot \sqrt{C/L}}\right) = R \cdot \left(1 + \frac{1}{1 + RC\omega_0}\right)$$
$$\Rightarrow \frac{|Z(\omega_0 + R/L)|}{|Z(\omega_0)|} = \sqrt{1 + \left(1 + \frac{1}{1 + RC\omega_0}\right)^2}.$$

For $\omega = \omega_0 - R/L$ we get

$$\frac{|Z(\omega_0 - R/L)|}{|Z(\omega_0)|} = \sqrt{1 + \left(1 + \frac{1}{1 - RC\omega_0}\right)^2}.$$

Note the asymmetry. The function $Z(\omega)$ is **not** symmetric around $\omega = \omega_0$.The effective power is according to (6.10)

$$\langle P_{\mathrm{el}}^{\mathrm{Wirk}} \rangle = \frac{1}{2}\frac{U_0^2 \cdot R}{|Z|^2}.$$

For $\omega = \omega_0 + R/L$ the power has decreases to the fraction

$$\frac{P(\omega_0 + R/L)}{P(\omega_0)} = \frac{1}{1 + \left(1 + \frac{1}{1 + RC\omega_0}\right)^2}$$

6.3 According to (6.15a, 6.15b) is

$$\omega_1 = \frac{\omega_0}{\sqrt{1-k}} = \frac{10^6}{\sqrt{1 - 0.05}} = 1.0266 \times 10^6\mathrm{s}^{-1},$$
$$\omega_2 = \frac{\omega_0}{\sqrt{1+k}} = \frac{10^6}{\sqrt{1 + 0.05}} = 0.9759 \times 10^6\mathrm{s}^{-1}.$$

The upper frequency ω_1 is 26 kHz above the resonance frequency ω_0, the lower frequency ω_2 24.1 kHz below ω_0. This shows that the resonance curve is not exactly symmetric. For the frequency $\nu = \omega/2\pi$ is ν_1 about 4.1 kHz above and ν_2 about 3.9 kHz below the center frequency ν_0.

6.4 The velocity of the electron is

$$v = \sqrt{2E_{\mathrm{kin}}/m} = \sqrt{27.2 \cdot 1.6 \times 10^{-19}/9.1 \times 10^{-31}}\,\mathrm{m/s} = 2.186 \times 10^6\,\mathrm{m/s}.$$

Its centrifugal acceleration on a circular path is

$$a = \frac{v^2}{r} = \frac{2.186^2 \times 10^{12}}{5.3 \times 10^{-11}}\frac{\mathrm{m}}{\mathrm{s}^2} = 9.19^{22}\,\mathrm{m/s}^2.$$

The radiated power is for a classical treatment (non-relativistic)

$$\bar{P} = \frac{e^2 a^2}{6\pi\varepsilon_0 c^3}.$$

This is identical to (6.38) when we set

$$a_x = d_0\omega^2 \cos\omega t$$
$$a_y = d_0\omega^2 \sin\omega t$$

The presence of two polarization-directions explains the difference to (6.38) by a factor 2.
Inserting the numerical values gives

$$\bar{P} = 4.6 \times 10^{-8}\,\mathrm{W}.$$

(a) The revolution period of the electron is

$$T = \frac{2\pi r}{\upsilon} = 1.5 \times 10^{-16}\mathrm{s}.$$

The energy radiated per circulation is

$$T \cdot \overline{\frac{\mathrm{dW}}{\mathrm{d}t}} = 1.5 \times 10^{-16} \cdot 4.6 \times 10^{-8}\,\mathrm{Ws} = 7 \times 10^{-24}\,\mathrm{Ws} = 4\,\mu\mathrm{eV}.$$

(b) The energy radiated per sec would be 44.6×10^{-8} Ws = 290 GeV

(c) When the electron loses energy by radiation it would follow in a classical model on a spiral path and finally crash into the nucleus. The quantitative calculation is as follows:
The total energy of the electron is the sum $W = E_{\mathrm{kin}} + E_{\mathrm{pot}}$.
On a stable circle with radius r around the nucleus the centripetal force is equal to the Coulomb-force:

$$\frac{m\upsilon^2}{r} = \frac{e^2}{4\pi\varepsilon_0 r^2}$$

Therefore the kinetic energy is

$$\Rightarrow E_{\mathrm{kin}} = \frac{m}{2}\upsilon^2 = \frac{1}{2}\frac{e^2}{4\pi\varepsilon_0 r} = \frac{1}{2}E_{\mathrm{pot}}$$

$$\Rightarrow W = +\frac{1}{2}E_{\mathrm{pot}} = -\frac{e^2}{8\pi\varepsilon_0 r},$$

$$\frac{\mathrm{d}W}{\mathrm{d}r} = +\frac{e^2}{8\pi\varepsilon_0 r^2} \Rightarrow \frac{\mathrm{dW}}{\mathrm{d}t} = \frac{e^2}{8\pi\varepsilon_0 r^2}\frac{\mathrm{d}r}{\mathrm{d}t}.$$

This is the mechanical power which is gained when the radius of the electron path decreases. It must be equal to the energy radiated away:

$$\frac{\mathrm{d}W}{\mathrm{d}t} = -\frac{e^2 a^2}{6\pi\varepsilon_0 c^3}$$

(the negative sign indicates that the energy of the electrons decreases).The acceleration a is

$$a = \frac{\upsilon^2}{r} = \frac{e^2}{4\pi\varepsilon_0 r^2 m}.$$

This shows that the radiation power depends on the radius r of the electron path. We obtain

$$-\frac{e^2}{6\pi\varepsilon_0 c^3}\cdot\left(\frac{e^2}{4\pi\varepsilon_0 m}\right)^2\cdot\frac{1}{r^4} = \left(\frac{\mathrm{d}W}{\mathrm{d}t}\right)_{\mathrm{em}}(r) \overset{!}{=} \frac{\mathrm{d}W}{\mathrm{d}t}\frac{\mathrm{d}r}{\mathrm{d}t} = \frac{e^2}{8\pi\varepsilon_0 r^2}\frac{\mathrm{d}r}{\mathrm{d}t}$$

$$\Rightarrow -r^2\mathrm{d}r = \frac{4}{3c^3}\left(\frac{e^2}{4\pi\varepsilon_0 m}\right)\mathrm{d}t$$

Integration from $r = a_0$ to $r = 0$ gives

$$a^3 = \frac{4}{c^3}\left(\frac{e^2}{4\pi\varepsilon_0 m}\right)\Delta t,$$

where $a_0 = 5.3 \times 10^{-11}$ m. The time it takes for the electron to reach the nucleus is

$$\Delta t \approx 1.6 \times 10^{-11}\mathrm{s}.$$

Note Experiments show, however, that the hydrogen atom in its lowest energy state is stable, i.e. the electron does not spiral into the nucleus. This can be only explained within the framework of quantum theory (see Vol. 3). In energetically higher states the atom radiates in deed energy in form of photons, which brings the atom back into the lowest stable energy state.

6.5 On a circular path with the radius R perpendicular to the magnetic field the centripetal force is equal to the Lorentz force.

$$\frac{m\cdot\upsilon^2}{R} = q\cdot\upsilon\cdot B \Rightarrow a = \frac{\upsilon^2}{R} = \frac{q}{m}\upsilon\cdot B.$$

The radiation power per sec is

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{q^2a^2}{6\pi\varepsilon_0 c^3} = \frac{q^4\upsilon^2 B^2}{6\pi\varepsilon_0 m^2 c^3}$$

$$= \frac{\mathrm{d}}{\mathrm{d}t}E_{\mathrm{kin}} = m\cdot\upsilon\cdot\frac{\mathrm{d}\upsilon}{\mathrm{d}t}$$

$$\Rightarrow \frac{\mathrm{d}\upsilon}{\mathrm{d}t} = \frac{q^4\upsilon\cdot B^2}{6\pi\varepsilon_0 m^2 c^3},$$

Here the change dv/dt of the amount of the velocity is assumed to be small compared to the change of the direction of v.
From the first equation we get the radius R of the circular path

$$R = \frac{m \cdot \upsilon}{q \cdot B}$$
$$\Rightarrow \frac{\mathrm{d}R}{\mathrm{d}t} = \frac{m}{q \cdot B}\frac{\mathrm{d}\upsilon}{\mathrm{d}t} = \frac{q^3 \cdot \upsilon B}{6\pi\varepsilon_0 m^2 c^3}$$
$$= \frac{\mathrm{d}W}{\mathrm{d}t} \cdot \frac{1}{q \cdot \upsilon \cdot B}.$$

6.6 (a, b) The accelerating force is

$$F = q \cdot E$$
$$\Rightarrow a = \frac{q}{m}E \Rightarrow |a| = a = \frac{q}{m} \cdot \frac{U}{d}.$$

Numerical values: $q = +1.6 \times 10^{-19}$ As, $m = 1.67 \times 10^{-27}$kg, $U = 10^6$V, $d = 3$ m, $\Rightarrow a = 3.2 \times 10^{13}$ m/s^2. The radiated power is then

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \frac{q^2 a^2}{6\pi\varepsilon_0 c^3} = 5.8 \times 10^{-27}\,\mathrm{W},$$

This is very small compared to the radiated power in the foregoing problem. The time needed for passing the acceleration length d can be obtained from

$$d = \frac{1}{2}at^2$$

which gives

$$t = \sqrt{\frac{2d}{a}} = \sqrt{\frac{6}{3.2 \times 10^{13}}}\,\mathrm{s} = 4.3 \times 10^{-7}\mathrm{s}.$$

During this time a proton looses the energy

$$\Delta W = \mathrm{d}W/\mathrm{d}t = 5.8 \times 10^{-27} \cdot 4.3 \times 10^{-7}\,\mathrm{Ws}$$
$$= 2.5 \times 10^{-33}\,\mathrm{Ws}.$$

This corresponds to the fraction

$$\eta = \frac{2.5 \times 10^{-33}}{1.6 \times 10^{-19} \times 10^6} = 1.5 \times 10^{-20}$$

of its acceleration energy and can be therefore neglected. On the circular path the acceleration is

$$a = \frac{\upsilon^2}{R} = \frac{2E_{\mathrm{kin}}}{m \cdot R}$$
$$= \frac{2 \times 10^6 \cdot 1.6 \times 10^{-19}}{1.67 \times 10^{-27} \cdot 3/2\pi}\,\frac{\mathrm{m}}{\mathrm{s}^2} = 4 \times 10^{14}\,\mathrm{m/s}^2.$$

The acceleration is here 12.5 times larger and the radiated power $\mathrm{d}W/\mathrm{d}t$, which is proportional to the square a^2 of the acceleration, is therefore 156 times larger.

6.7 The intensity I of the wave is equal to the energy-flux-density at a distance of 1 m from the source.

$$I = |S| = \frac{P_{\mathrm{em}}}{4\pi r^2} = \frac{10^4\mathrm{W}}{4\pi \cdot 1\,\mathrm{m}^2} = 8 \times 10^2\,\mathrm{W/m}^2.$$

The electric field strength E is according to (6.36a)

$$E = \sqrt{S/(\varepsilon_0 \cdot c)} = 5.5 \times 10^2\,\mathrm{V/m}.$$

The magnetic field strength is

$$B = \frac{1}{c}E = 1.83 \times 10^{-6}\,\frac{\mathrm{Vs}}{\mathrm{m}^2} = 1.83\,\mu\mathrm{T}.$$

6.8 The energy flux density is equal to the pointing vector S

$$S = \frac{\bar{P}_{\mathrm{em}}}{4\pi r^2 \cdot \Delta\Omega} \Rightarrow \bar{P}_{\mathrm{em}} = 4\pi r^2 \times 10^{-2} \cdot S.$$

S is defined as

$$S = \varepsilon_0 c E^2 = 8.85 \times 10^{-12} \cdot 3 \times 10^8 \times 10^2\,\mathrm{W/m}^2$$
$$= 0.26\,\mathrm{W/m}^2$$

We therefore get the average emitted power as

$$\bar{P}_{\mathrm{em}} = 3.27 \times 10^4\mathrm{W}.$$

From (6.38) it follows with $q = N \cdot e$

$$\bar{P}_{\mathrm{em}} = \frac{N^2 e^2 \cdot 16\pi^4 \nu^4 d_0^2}{12\pi\varepsilon_0 c^3}$$
$$\Rightarrow d_0 = \sqrt{\frac{3\varepsilon_0 \cdot c^3 \cdot \bar{P}_{\mathrm{em}}}{N^2 e^2 \cdot 4\pi^3 \nu^4}}.$$

Inserting the numerical values $N = 10^{28} \times 10^{-4} \times 10 = 10^{25}$, $\nu = 10^7\ \mathrm{s}^{-1}$, $e = 1.6 \times 10^{-19}$ C gives

$$\Rightarrow d_0 = 2.7 \times 10^{-12}\,\mathrm{m}.$$

This illustrates that the oscillation amplitude of the oscillating electrons is very small.

6.9 (a) The solar constant gives the energy flux density at the upper edge of the earth atmosphere where the absorption of the sunlight can be still neglected. From

$$S = \varepsilon_0 \cdot c \cdot E^2$$

we can derive the electric field of the sun radiation at the upper edge of our atmosphere

$$E = \sqrt{\frac{S}{\varepsilon_0 \cdot c}} = \sqrt{\frac{1.4 \times 10^3}{8.85 \times 10^{-12} \cdot 3 \times 10^8}}\,\frac{\mathrm{V}}{\mathrm{m}}$$
$$= 7.26 \times 10^2\,\mathrm{V/m}$$

The magnetic field is then

$$\Rightarrow B = \frac{1}{c} \cdot E = \frac{7.26 \times 10^2 \,\mathrm{V\,s}}{3 \times 10^8 \,\mathrm{m^2}} = 2.4 \times 10^{-6}\,\mathrm{T}.$$

(b) the distance from the center of the sun to the earth is $r = 1.5 \times 10^{11}$ m.
The total power radiated by the sum is then

$$\bar{P}_{\mathrm{em}} = 4\pi r^2 \cdot S = 1.4 \times 10^3 \cdot 4\pi \cdot 1.5^2 \times 10^{22}\,\mathrm{W} = 4 \times 10^{26}\,\mathrm{W}.$$

(c) The energy flux density at the surface of the sun is

$$S_\odot = \frac{\bar{P}_{\mathrm{em}}}{4\pi R_\odot^2} = \frac{4 \times 10^{26}}{4\pi \cdot 6.96^2 \times 10^{16}} = 6.57 \times 10^7\,\mathrm{W/m^2}$$

$$\Rightarrow E = \sqrt{\frac{S}{\varepsilon_0 c}} = 1.57 \times 10^5\,\mathrm{V/m}.$$

6.10 As in Problem 6.9 is

$$S = \frac{\bar{P}_{\mathrm{em}}}{4\pi r^2}, \quad E = \sqrt{\frac{S}{\varepsilon_0 c}}.$$

With $r = 1$ m and $\bar{P}_{\mathrm{em}} = 70$W is $E = 45$ V/m.
In order to reach the same electric field strength as that of the sun radiation one must increase the energy flux density by a factor $k = (726/45)^2 = 260$. This implies that also the power must be larger by a factor of 260, i.e. it must be 26 kW.
Note, however

(a) that absorption and scattering in the earth atmosphere decreases the radiation flux density of the sun by a factor 0.5–0.6
(b) Only a fraction of the sun radiation falls into the visible range (see the spectral curve of the sun radiation (Vol. 3, Chap. 2) with a temperature of the sun surface of about 5800 K.

Chapter 7

7.1 From the equation rot $B = \varepsilon_0\mu_0 \partial E/\partial t$ it follows

$$\mathbf{rot\,rot}\,\boldsymbol{B} = \varepsilon_0\mu_0 \frac{\partial}{\partial t}(\mathbf{rot}\,\boldsymbol{E}) = -\varepsilon_0\mu_0 \frac{\partial^2 \boldsymbol{B}}{\partial t^2},$$

$$\mathbf{rot\,rot}\,B = \mathbf{grad}(\mathrm{div}\,\boldsymbol{B}) - \mathbf{divgrad}\,\boldsymbol{B} = -\Delta\boldsymbol{B},$$

because **div B = 0**. We therefore get

$$\Delta B = \varepsilon_0\mu_0 \frac{\partial^2 B}{\partial t^2} = \frac{1}{c^2}\frac{\partial^2 B}{\partial t^2}.$$

7.2 A plane wave propagating into the z-direction is described by

$$E = E_0 \cdot \mathrm{e}^{\mathrm{i}(\omega t - k \cdot r)}.$$

For $\mathbf{k} \cdot \mathbf{r} = 0$ the phase $\varphi = \omega \cdot t_0 - \mathbf{k} \cdot \mathbf{r}$ has for a given time t_0 for all positions r the same value. This means: The geometrical position for all vectors r with $\mathbf{k} \cdot \mathbf{r} =$ const represents the *phase surface* (Fig. A.24). From $\mathbf{k} \cdot \mathbf{r_1} = \mathbf{k} \cdot \mathbf{r_2} =$ const. $\rightarrow \mathbf{k} \cdot (\mathbf{r_1} - \mathbf{r_2}) = 0 \rightarrow \mathbf{k} \perp (\mathbf{r_1} - \mathbf{r_2})$. The vector $\mathbf{r_1} - \mathbf{r_2}$ is a vector in the plane $\perp \mathbf{k}$. This implies that the plane $\perp \mathbf{k}$ is a phase plane.

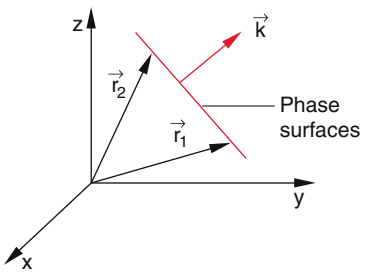


Fig. A.24 Illustration of solution 7.2

7.3 With $E = a_1E_1 + a_2E_2 \rightarrow$

$$I = \varepsilon_0 c E^2 = \varepsilon_0 c\left[a_1^2E_1^2 + a_2^2E_2^2 + 2a_1a_2E_1 \cdot E_2\right].$$

With $E_i = E_0 \cdot \cos(\omega t + \varphi_i)$ we obtain

$$\bar{I} = \varepsilon_0 c\overline{E^2} = \frac{1}{2}\varepsilon_0 c\left[a_1^2E_{01}^2 + a_2^2E_2^2 + 2a_1a_2E_{10}E_{20\cos(\varphi_1-\varphi_2)}\right] = \bar{I}_1 + \bar{I}_2 + 2 \cdot \sqrt{I_1I_2} \cdot \overline{\cos(\varphi_1 - \varphi_2)}.$$

For incoherent light the phase-differences $\Delta\phi = \phi_1(t) - \varphi_2(t)$ fluctuate randomly and therefore is the mean value $\overline{\cos\varphi_1 - \varphi_2} = 0$. In this case is the total intensity equal to the sum of the intensities of the partial waves. For coherent light this is no longer true.

7.4 The representation of a circular polarized wave is

$$cE = A \cdot \mathrm{e}^{\mathrm{i}(\omega t - kz)} \quad \text{mit} \quad \mathrm{A} = \mathrm{A}_0(\hat{x} \pm \mathrm{i}\hat{y})$$
$$\sigma^+ - \text{Licht}: \mathrm{A} = \mathrm{A}_0(\hat{x} + \mathrm{i}\hat{y})$$
$$\sigma^- - \text{Licht}: \mathrm{A} = \mathrm{A}_0(\hat{x} - \mathrm{i}\hat{y})$$
$$E^+ + E^- = 2A_0\hat{x}\mathrm{e}^{\mathrm{i}(\omega t - kz)}$$

This is a linear-polarized wave with the **E**-vector pointing into the x-direction.

7.5 For stationary equilibrium the total input power must be equal to the total output power. This gives the equation

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \alpha \cdot I \cdot F \cdot \cos\gamma - c_W \cdot \frac{\mathrm{d}M}{\mathrm{d}t}(T - T_U) - \kappa(T - T_U) = 0.$$

where α is the fraction of the absorbed power. The mass of the water running per sec through the pipe is

$$\frac{\mathrm{d}M}{\mathrm{d}t} = \frac{\alpha \cdot I \cdot F \cdot \cos\gamma}{c_\mathrm{W}(T - T_U)} - \frac{\kappa}{c_\mathrm{W}}.$$

With the numerical values $\alpha = 0.8$; $I = 500\ \mathrm{W/m^2}$, $\cos\gamma = 0.94$; $c_\mathrm{w} = 4.18\ \mathrm{kJ/kg}$; $T - T_\mathrm{u} = 60\ \mathrm{K}$, $\kappa = 2\ \mathrm{W/K}$ we obtain

$$\begin{aligned}\frac{\mathrm{d}M}{\mathrm{d}t} &= \frac{0.8 \cdot 500 \cdot 4 \cdot 0.94}{4.18 \times 10^3 \cdot 60} - 0.48 \times 10^{-3}\,\mathrm{kg/s} \\ &= \left(6 \times 10^{-3} - 0.48 \times 10^{-3}\right)\mathrm{kg/s} \\ &\approx 5.5 \times 10^{-3}\,\mathrm{l/s} = 201/\mathrm{h}.\end{aligned}$$

The mean sun energy density incident in June per sunny day per $\mathrm{m^2}$ is about 6 kWh. This is sufficient to heat per day and $\mathrm{m^2}$ collector area 60 l water by 60 K (from 20 to 80 °C). if the heat losses can be neglected ($\kappa = 0$)

7.6 (a) We regard a capacitor with circular plates and areas $A = 4\pi r^2$ and the plate distance d. The charge Q is then

$$Q = C \cdot U = \varepsilon_0 \frac{A}{d} U = \varepsilon_0 \cdot A \cdot E$$

With $\mathbf{E} = \{0, 0, E\}$ the intensity is,

$$I = \frac{\mathrm{d}Q}{\mathrm{d}t} = \varepsilon_0 A \cdot \frac{\partial E}{\partial t}.$$

The magnetic field lines are circles around the z-axis

$$\begin{aligned}\oint B\,\mathrm{d}s = B(r) \cdot 2\pi r &= \mu_0 \cdot \frac{r^2}{R^2} I \\ \Rightarrow B(r) &= \frac{\mu_0 I}{2\pi R^2} r.\end{aligned}$$

(b) The Pointing vector is

$$S = \varepsilon_0 c^2 (E \times B).$$

it has only a radial component in a plane perpendicular to the z-axis. Its amount is

$$\begin{aligned}|S| &= \varepsilon_0 c^2 \cdot \frac{Q}{\varepsilon_0 A} \cdot \frac{\mu_0 I}{2\pi R^2} r \\ &= \frac{Q \cdot I \cdot r}{2\varepsilon_0 A^2} = \frac{r}{2\varepsilon_0 A^2} \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2} Q^2\right).\end{aligned}$$

(c) The energy flux passing per sec through the cylinder surface $2\pi \cdot r \cdot d$ is

$$\begin{aligned}\frac{\mathrm{d}W}{\mathrm{d}t} &= |S| \cdot 2\pi r \cdot d \\ &= \frac{\pi r^2 \cdot d}{\varepsilon_0 A^2} \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2} Q^2\right) \\ &= \frac{\pi r^2}{A} \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2} C \cdot U^2\right).\end{aligned}$$

This is the fraction of the energy $W = ½C \cdot U^2$ stored in the volume $\pi r^2 \cdot d$ of the capacitor which streams per second out of the capacitor.

7.7 The earth appears from the sun under the solid angle

$$\Omega_\mathrm{E} = \frac{\pi R_\mathrm{E}^2}{(1\,\mathrm{AE})^2}.$$

Mars appears under the angle

$$\Omega_\mathrm{M} = \frac{\pi R_\mathrm{M}^2}{(1.52\,\mathrm{AE})^2}.$$

We then get

$$\frac{S_\mathrm{M}}{S_\mathrm{E}} = \frac{R_\mathrm{M}^2}{R_\mathrm{E}^2 \cdot 1.52^2} = \frac{0.532^2}{1 \cdot 1.52^2} = 0.123,$$

where the radius of Mars is $R_\mathrm{M} = 0.532\ R_\mathrm{E}$ and the distance Sun-Mars is 1.52 AE.
The power reflected by Mars into the solid angle 2π is

$$S_\mathrm{MR} = 0.5 \cdot 0.123\, S_\mathrm{E}.$$

The solid angle of the earth seen from Mars at its closest approach is

$$\Omega_\mathrm{ME} = \frac{\pi R_\mathrm{E}^2}{(0.52\,\mathrm{AE})^2}.$$

The diffuse sun radiation reflected by Mars and received by the earth is

$$\frac{\mathrm{d}W_\mathrm{ME}}{\mathrm{d}t} = \frac{0.5 \cdot 0.123\, S_\mathrm{E} \cdot \pi R_\mathrm{E}^2}{(0.52\,\mathrm{AE})^2 \cdot 2\pi} = 1.9 \times 10^{-9}\, S_\mathrm{E}.$$

Mars therefore radiates to the earth at its closest approach only 1.9×10^{-9} times the radiation power which is directly received by the earth from the sun.

7.8 The maximum radiation power transmitted by the pupil of the eye is

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \left(800\,\mathrm{W/m^2}\right) \cdot \pi r^2 = 800\pi \times 10^{-6}\,\mathrm{W} = 2.5\,\mathrm{mW}.$$

The intensity on the retina is then

$$I = \frac{A_{\text{Pupille}}}{A_{\text{Netthaut}}} I_0 = 400 I_0 = 320\,\text{kW/m}^2.$$

This is sufficient to destroy the photo-receptors of the retina.

7.9 The weight force $m \cdot g$ has to be balanced by the light pressure. With the intensity I of the radiation incident into the z-direction the light pressure exerted onto the area A oriented under the angle ϑ against the z-direction is

$$dA_z = dA \cdot \cos\vartheta = 2\pi R^2 \cdot \sin\vartheta \cos\vartheta d\vartheta.$$

A circular stripe with the radius $a = R \cdot \sin\vartheta$ has the area $dA = 2\pi a \cdot R \cdot d\vartheta$ (Fig. A.25). The projection perpendicular to the incident light is

$$\frac{dp_e}{dt} = \frac{I}{c} dA_z$$

The momentum transfer by the reflected light is

$$\frac{dp_r}{dt} = \frac{I}{c} \cos(2\vartheta) dA_z.$$

The other components dp_x/dt and dp_y/dt cancel when integrating over the total stripe.
Integration over the lower half sphere gives

$$\begin{aligned}\frac{dp}{dt} &= \frac{I}{c} \int_0^{\pi/2} (1 + \cos 2\vartheta) dA_z \\ &= 2\pi R^2 \frac{I}{c} \int_0^{\pi/2} (1 + \cos 2\vartheta) \sin\vartheta \cos\vartheta d\vartheta \\ &= \pi R^2 \cdot I/c.\end{aligned}$$

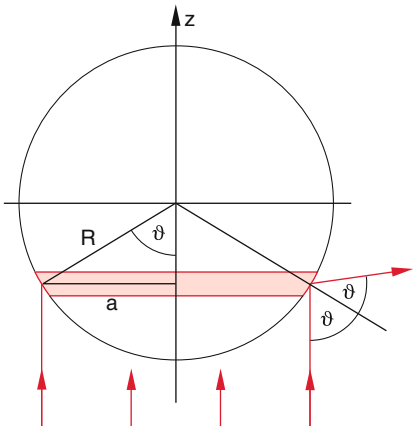


Fig. A.25 Illustration to solution 7.9

Only the first term in the integral gives a contribution because the integration over $\cos^2\vartheta \cdot \sin\vartheta \cdot \cos\vartheta$ gives zero.
The momentum transfer is therefore only caused by the incident light and not by the reflected light. It has half the amount as for a circular disc with the area $\pi \cdot R^2$ oriented perpendicular to the incident radiation because for the disc the reflected light contributes the same amount to the momentum transfer as the incident light, while for the sphere the reflected light transfers the momentum for $\vartheta < 45°$ into the $+z$-direction but for $\vartheta > 45°$ into the $-z$-direction. The two contributions therefore cancel each other.
Question: Could this result have been obtained immediately by intuition without the lengthy calculation?
The necessary intensity of the incident light is the for a mass density $\rho = m/V$ of the sphere

$$I = \frac{m \cdot g \cdot c}{\pi R^2} = \frac{4}{3} R \cdot \varrho \cdot g \cdot c.$$

The result applies for the absorbing as well as for the reflecting sphere.

7.10 For the arbitrary position of the light mill in Fig. A.26 the incident parallel light beam forms the angle α against the surfaces 1 and 3 and the angle $\beta = 90° - \alpha$ against the surfaces 2 and 4. The radiation pressure onto the reflecting surfaces 1 and 2 causes a clockwise torque the radiation pressure onto the absorbing surfaces 3 and 4 an anticlockwise torque. The area of each surface is $A = a^2$.
Radiation with the intensity I exerts according to (7.27) the force

$$dF = \frac{2I}{c} a \cdot ds \cdot \sin\alpha \cdot \hat{e}_x$$

onto the surface element $dA_1 = a \cdot dx$ of the surface 1. This force causes a torque $dD_1 = dF_1 \times s$ about the z-axis. With $y = s \cdot \sin\alpha$ the amount of the torque is

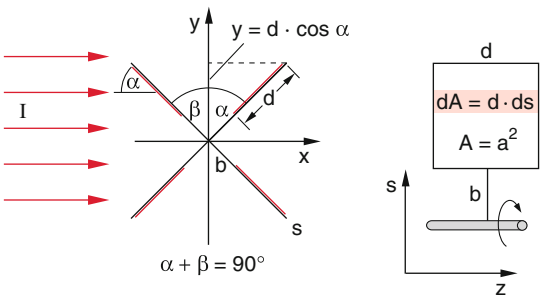


Fig. A.26 Illustration to solution 7.10a

$$
\begin{aligned}
\mathrm{d}D_1 &= \mathrm{d}F_1 \cdot s \cdot \sin\alpha = \frac{2I}{c} a \sin^2\alpha \cdot s\mathrm{d}s \\
&= \frac{2I}{c} ay \cdot \mathrm{d}y \\
\Rightarrow D_1 &= \frac{2I}{c} \cdot a \cdot \int_{b\cdot\sin\alpha}^{(b+a)\sin\alpha} y\mathrm{d}y \\
&= \frac{I}{c} a \sin^2\alpha\left(a^2 + 2ba\right).
\end{aligned}
$$

The torque of the surface 2 is accordingly

$$
D_2 = \frac{2I}{c} a \int_{y_1}^{y_2} y\,\mathrm{d}y = \frac{I}{c} a\left[y_2^2 - y_1^2\right]
$$

with $y = s \cdot \cos\alpha$. The surface 2 is partly obscured by the surface 1. Therefore only a part of surface 2 is illuminated. For $\alpha < 45°$ this is the part from $y_1 = (a + b) \cdot \cos\beta$ until $y_2 = (a + b) \cdot \sin\beta$. Because $\cos\beta = \sin\alpha$ we can write: $y_1 = (a + b) \sin\alpha$ and $y_2 = (a + b) \cos\alpha$. This gives

$$
\begin{aligned}
D_2 &= \frac{I}{c} a(a+b)^2\left[\sin^2\alpha - \cos^2\alpha\right] \\
&= \frac{I}{c} a(a+b)^2\left[1 - 2\cos^2\alpha\right].
\end{aligned}
$$

For $\alpha \geq 45°$ the illuminated part reaches from $y_1 = b \cdot \cos\alpha$ until $y_2 = b \cdot \sin\alpha$, which gives for the torque. The torque D_3 can be obtained in a similar way as D_1 by replacing α by $\beta = 90° - \alpha$ and taking into account that the absorbing surface experiences only ½ of momentum transfer as the reflecting surface.

$$
\begin{aligned}
\Rightarrow D_3 &= -\frac{I}{2c} a \cdot \cos^2\alpha\left[a^2 + 2ba\right], \\
D_4 &= -\frac{I}{2c} a(a+b)^2\left(1 - 2\sin^2\alpha\right) \\
&\quad \text{for} \quad \alpha \leq 45^\circ, \\
D_4 &= \frac{I}{2c} ab^2\left(1 - 2\sin^2\alpha\right) \quad \text{for} \quad \alpha \geq 45^\circ.
\end{aligned}
$$

The total torque is $D = D_1 + D_2 + D_3 + D_4$. With $b = 1$ cm, $a = 2$ cm, $I = 10^4$ W/m^2 one obtains:

$$
\begin{aligned}
D_1 &= \frac{I}{c} \cdot \sin^2\alpha \cdot 16 \times 10^{-6}\,\mathrm{Nm} \\
&= 5.3 \times 10^{-10} \cdot \sin^2\alpha\mathrm{Nm}, \\
D_2 &= 6 \times 10^{-10}\left[\sin^2\alpha - \cos^2\alpha\right], \\
D_3 &= -2.67 \times 10^{-10}\cos^2\alpha, \\
D_4 &= -3 \times 10^{-10}\left[\cos^2\alpha - \sin^2\alpha\right], \\
\Rightarrow D &= 14.3 \times 10^{-10}\sin^2\alpha \\
&\quad - 11.67 \times 10^{-10}\sin^2\alpha \quad \text{for} \quad \alpha \leq 45^\circ.
\end{aligned}
$$

This is by far too small to turn the light mill. Therefore the mill will not turn in vacuum.

(b) We assume, that the incident light increases the temperature only of the absorbing black surface but not that of the reflecting, non-absorbing surface. The temperature increase ΔT of the absorbing surface can be obtained from

$$
\Delta T = \frac{1}{C_\mathrm{W}}\left(I \cdot \Delta A - \frac{\mathrm{d}W}{\mathrm{d}t} \cdot \Delta T\right),
$$

where C_w is the heat capacity of one plate of the light mill, ΔA is the illuminated area and $(\mathrm{d}W/\mathrm{d}t) \cdot \Delta\mathrm{T}$ the power taken away from the plate by collisions with argon atoms.

$$
\Rightarrow \Delta T = \frac{I \cdot \Delta A}{C_\mathrm{W} + \mathrm{d}W/\mathrm{d}t}
$$

An atom has the kinetic energy E_kin = ½mv^2 = 3/2 $k \cdot T$ before the collision with the plate and 3/2 $k \cdot (T + \Delta T)$ after the collision. The thermal power taken away from the plate is then (see Vol. 1, Sect. 7.5.3)

$$
\frac{\mathrm{d}W}{\mathrm{d}t} \cdot \Delta T = \frac{n}{4} \cdot \frac{3}{2} k\Delta T \cdot \overline{v} \cdot A
$$

(n = atom number density). With $n = 3 \times 10^{16}$/cm^3, $A = 4$ cm^2, $v = 5 \times 10^4$ cm/s, $k = 1.38 \times 10^{-23}$ J/K we get

$$
\frac{\mathrm{d}W}{\mathrm{d}t} = 0.031\mathrm{W}.
$$

Each surface is illuminated during ¼ of the circulation period. With $\overline{\sin^2\alpha} = 1/2$ the mean illumination power is

$$
\begin{aligned}
\overline{I \cdot \Delta A} &= \frac{1}{2} \cdot \frac{1}{4} I \cdot A \\
&= \frac{1}{8} \times 10^4\,\mathrm{W/m^2} \cdot 4 \times 10^{-4}\mathrm{m^2} = 0.5\,\mathrm{W} \\
\Rightarrow \Delta T &= \frac{0.5}{0.13} \approx 4\,\mathrm{K}.
\end{aligned}
$$

The black surface is heated up by $\Delta T = 4$ K.
The momentum transfer from the colliding atoms onto the surface is

$$
\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{n \cdot m}{4} \cdot \left[\overline{v}(T + \Delta T) - \overline{v}(T)\right] \cdot A\overline{v}
$$

(see Vol. 1, Eq. (7.47) where $m = 40 \cdot 1.67 \times 10^{-27}$ kg is the mass of one argon atom and v its mean velocity.

$$
\begin{aligned}
\overline{v} &= \sqrt{\frac{8kT}{\pi \cdot m}} \\
\Rightarrow \frac{\mathrm{d}p}{\mathrm{d}t} &= \frac{n \cdot m}{4} A \frac{8k}{\pi \cdot m}\left(\sqrt{T(T + \Delta T)} - T\right) \\
&= \frac{3}{4} \times 10^{22} \cdot 4 \times 10^{-4} \cdot \frac{8}{\pi} \\
&\quad \times 1.38 \times 10^{-23} \cdot \left(\sqrt{300 \cdot 304} - 300\right), \\
F &= 5.3 \times 10^{-5}\,\mathrm{N}.
\end{aligned}
$$

The mean torque is then (similar to solution 7.10a)

$$D = F \cdot (b + a/2) \approx 10^{-6}\text{N} \cdot \text{m},$$

This is more than 3 orders of magnitude larger the torque caused by photon recoil.

7.11 The power radiated by the antenna has rotational symmetry around the antenna axis and is proportional to $\sin^2\vartheta$. The power radiated into the solid angle

$$\mathrm{d}P = P_0 \cdot \sin^2\vartheta\,\mathrm{d}\vartheta$$

is

$$\begin{aligned}\mathrm{d}\Omega &= \frac{1}{r^2} \cdot r\,\mathrm{d}\alpha \cdot r \cdot \sin\alpha\,\mathrm{d}\varphi \\ &= \sin\alpha\,\mathrm{d}\alpha\,\mathrm{d}\varphi.\end{aligned}$$

Integration over all angles φ (the parabolic mirror has rotational symmetry around the x-axis) gives the power radiated into the angular range from ϑ to $\vartheta + \mathrm{d}\vartheta$

$$\begin{aligned}\mathrm{d}P &= P_0 \cdot \sin^2\vartheta\,\sin\alpha \cdot 2\pi \cdot \mathrm{d}\alpha \quad (\alpha = 90^\circ - \vartheta) \\ &= -P_0 \sin^2\vartheta\,\cos\vartheta \cdot 2\pi \cdot \mathrm{d}\vartheta, \\ P &= -2\pi P_0 \int_{\vartheta=90^\circ} \sin^2\vartheta\,\cos\vartheta\mathrm{d}\vartheta \\ &= \frac{2\pi}{3} P_0 \sin^3\vartheta|_{\vartheta_{\min}}^{\pi/2} \\ &= \frac{2\pi}{3} P_0 \left(1 - \sin^3\vartheta_{\min}\right).\end{aligned}$$

integration over ϑ gives The angle $\vartheta_{\min}$ can be obtained from cos $\vartheta = y/r$

$$\cos\vartheta_{\min} = \frac{D/2}{\sqrt{y^2 + (f - x)^2}}.$$

With $y^2 = 4 \cdot f \cdot$ x (equation for the parabolic surface) it follows:

$$\begin{aligned}\cos\vartheta_{\min} &= \frac{D/2}{f + x}, \\ x &= \frac{D^2}{16f}\end{aligned}$$

$$\begin{aligned}&\Rightarrow \cos\vartheta_{\min} = \frac{D/2}{f + D^2/16f} = \frac{8Df}{D^2 + 16f^2}, \\ &P = \tfrac{2\pi}{3} P_0 \left(1 - \sin^3\left[\arccos\frac{8Df}{D^2 + 16f^2}\right]\right).\end{aligned}$$

7.12 $v_{\mathrm{G}} = 1/3c = c^2/v_{\mathrm{Ph}}$

$$\begin{aligned}&\Rightarrow v_{\mathrm{Ph}} = 3c = \frac{c}{\sqrt{1 - \dfrac{n^2\pi^2c^2}{a^2\omega^2}}} \\ &\Rightarrow \frac{n^2\pi^2c^2}{a^2\omega^2} = \frac{8}{9} \Rightarrow \lambda^2 = \frac{4a^2}{n^2} \cdot \frac{8}{9}.\end{aligned}$$

The maximum wavelength is obtained for n = 1

$$\Rightarrow \lambda_{\max} = \frac{2a}{3} \cdot \sqrt{8}\,\text{cm} = 5.66\,\text{cm}.$$

7.13 $U = I \cdot R$ = 3 × 10 V = 30 V. If the current streams into the z-direction the electric field has only a z-component. Its amount is

$$E = \frac{U}{L} = \frac{30}{100}\frac{\text{V}}{\text{m}} = 0.3\,\text{V/m}$$

The magnetic field on the surface of the wire is

$$B = \frac{\mu_0 I}{2\pi r_0} = \frac{4\pi \times 10^{-7} \times 10}{2\pi \cdot 3 \times 10^{-3}}\text{T} = 0.67\,\text{mT}.$$

The Poynting vector points into the radial direction towards the wire axis. Its amount is

$$S = \frac{1}{\mu_0} E \cdot B = \frac{I \cdot U}{2\pi r_0 \cdot L}.$$

S gives the energy floux per sec and surface area A = 1 cm^2. The total power streaming into the wire is then for a wire surface $A = 2P \cdot r_0 \cdot L$

$$\frac{\mathrm{d}W}{\mathrm{d}t} = U \cdot I = I^2 \cdot R,$$

This is equal to the energy loss $I^2 \cdot R$ due to the Ohmic resistance R.

7.14 The photon recoil per sec is according to (7.26)

$$\frac{\mathrm{d}p}{\mathrm{d}t} = F_{\mathrm{R}} = \varepsilon_0 E^2 \cdot A = m \cdot a.$$

with $I = \varepsilon_0 \cdot c \cdot E^2$ we get

$$I = \frac{c \cdot m \cdot a}{A}.$$

If the acceleration $a = 10^{-5}$ m/s^2 should be reached for a mass of $m = 10^3$ kg and an area of $A = 10^{-2}$ m^2 the intensity must be at least

$$I = 3 \times 10^8\,\text{W/m}^2$$

The radiation power of the lamp must be then

$$P_{\text{Licht}} = I \cdot A = 3 \times 10^6\,\text{W}$$

Remark More realistic are space ships with huge solar sails which can use the light pressure of the sun radiation, for instance for a journey to Mars. With an area $A = 10^4\ \text{m}^2$ and the sun intensity $I = 10^3\ \text{W/m}^2$ one obtain the acceleration

$$a = \frac{2I \cdot A}{m \cdot c} = 6.6 \times 10^{-5}\text{m/s}^2$$

without any energy consumption of the energy storage in the space ship.

7.15 As has been discussed in Sect. 1.3.4 the electric field between the outer and the inner conducting cylinder of the coaxial waveguide is

$$E = \frac{\lambda}{2\pi\varepsilon_0 r}\hat{r} \quad \text{for} \quad \text{a} \le \text{r} \le \text{b}.$$

for $a \le r \le b$.
The voltage between inner and outer cylinder is then

$$U = \int_a^b E\,\text{d}r = \frac{\lambda}{2\pi\varepsilon_0}\ln(b/a),$$

where $\lambda = Q/l$ is the charge per unit length. The capacity per unit length is then

$$\hat{C} = \frac{\lambda}{U} = \frac{2\pi\varepsilon_0}{\ln(b/a)}.$$

The inductance per unit length is according to Problem 4.2

$$\hat{L} = \frac{\mu_0}{2\pi}\ln\frac{b}{a} \Rightarrow \hat{C} \cdot \hat{L} = \varepsilon_0 \cdot \mu_0 = \frac{1}{c^2},$$

Note that L is independent of the geometry of the coaxial cable.
The wave impedance of the coaxial waveguide is

$$Z_0 = \sqrt{\hat{L}/\hat{C}} = \frac{1}{2\pi}\sqrt{\frac{\mu_0}{\varepsilon_0}}\ln\frac{b}{a}$$

$$= \frac{\mu_0 \cdot c}{2\pi}\ln\frac{b}{a}$$

$$\Rightarrow b = a \cdot \exp\left[\frac{2\pi Z_0}{\mu_0 \cdot c}\right].$$

For $Z_0 = 100\,\Omega, a = 10^{-3}\,\text{m}$ follows $b = 10^{-3} \cdot \text{e}^{10/6}\,\text{m} = 5.3\,\text{mm}$.

Chapter 8

8.1 At atmospheric pressure the molecular number density is $N \approx 2.5 \times 10^{25}/\text{m}^3$. For a wavelength $\lambda = 500$ nm we get $\omega = 3.77 \times 10^{15} s^{-1}$.
The electron mass is $m = 9.1 \times 10^{-31}$ kg.

$$\omega_0^2 - \omega^2 = \left(1 - 0.377^2\right) \times 10^{32}$$
$$= 0.86 \times 10^{32} \gg \gamma \cdot \omega$$
$$n = 1 + \frac{2.5 \times 10^{25} \cdot 1.6^2 \times 10^{-38}}{2 \cdot 8.8 \times 10^{-12} \cdot 9.1 \times 10^{-31} \cdot 0.86 \times 10^{32}}$$
$$= 1 + 4.6 \times 10^{-4}$$

The comparison with Table 8.1 shows that the experimental value is $(n-1)_{\text{ex}} = 2.79 \times 10^{-4}$.
This means that by the comparison with Eq. 8.23 that the oscillator strength for the strongest transition starting from the ground state is $f_1 \approx 2.79/4.6 = 0.6$. This means that the molecules show on their ground state absorption at $\lambda = 190$ nm an absorptivity which corresponds to 60% of a classical oscillator (Fig. A.27).

8.2 For the angles ∠ir between incident and reflected light beam and ∠ig between incident and refracted beam we get

$\angle\text{ir} = 2\alpha$, $\angle\text{ig} = 180° + \beta - \alpha$ (Fig. A.28)

$$\Rightarrow 2\alpha = 180° - \alpha + \beta,$$
$$\Rightarrow 3\alpha = 180° + \beta,$$
$$\Rightarrow \sin 3\alpha = \sin(180° + \beta) = -\sin\beta$$
$$= -\frac{1}{n}\sin\alpha$$
$$\Rightarrow \frac{1}{n} = -\frac{\sin 3\alpha}{\sin\alpha} = \frac{4\sin^3\alpha - 3\sin\alpha}{\sin\alpha}$$
$$= 4\sin^2\alpha - 3$$
$$\Rightarrow \sin\alpha = \sqrt{\left(3 + \frac{1}{n}\right)/4}.$$

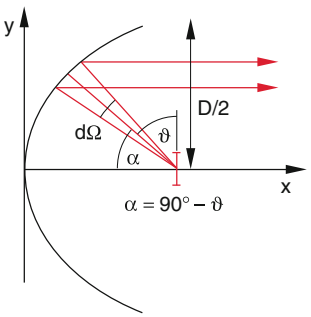


Fig. A.27 Illustration to solution 7.11

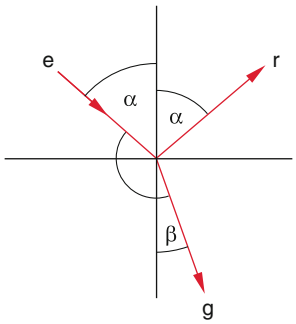


Fig. A.28 Illustration to solution 8.2

For $n = 1.5$ we get

$$\sin\alpha = \sqrt{0.91666} \approx 0.957$$
$$\Rightarrow \alpha = 73.3^\circ.$$

8.3 We assume that the incident wave propagates parallel to the z-direction and its E-vector is parallel to the x-direction. The scattered radiation is observed in the y-direction (Fig. A.29). The atoms 5–8 are later excited than the atoms 1–4. The phaseshift is

$$\Delta\varphi = \frac{d}{\lambda}\cdot 2\pi = \frac{1}{3}\cdot 2\pi = \frac{2}{5}\pi.$$

The light emitted by the atoms 1, 2, 5 and 6 is detected with a phase shift $\Delta\varphi$ against the light detected from the atoms 4, 3, 7, 8. If we set the phase of the light scattered by the atoms 3 and 4 as $\varphi = 0$, the wave scattered by the atoms 1, 2, 7 and 8 the phase shift $\Delta\varphi$, that emitted by the atoms 5 and 6 the phase shift $23\Delta\varphi$. The total amplitude of the scattered light is then

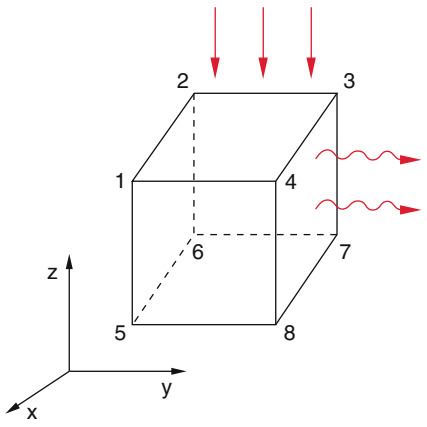


Fig. A.29 Illustration to solution 8.3

$$\begin{aligned} A &= A_0\cdot e^{i\omega t}\left(2+4\cdot e^{i\Delta\varphi}+2\cdot e^{2i\Delta\varphi}\right)\\ &= A_0\cdot e^{i(\omega t+\Delta\varphi)}\cdot\left(4+4\cdot\frac{e^{i\Delta\varphi}+e^{-i\Delta\varphi}}{2}\right)\\ &= A_0\cdot e^{i(\omega t+\Delta\varphi)}(4+4\cos\Delta\varphi)\\ &\Rightarrow P = P_0\cdot 16(1+\cos\Delta\varphi)^2\\ &= 16\,P_0\cdot 4\cos^4(\Delta\varphi/2), \end{aligned}$$

$$\Delta\phi = 2\pi\cdot(1/\lambda) = 2\pi\cdot(100/500) = (2/5)\pi$$
$$\rightarrow \cos^4(\Delta\phi/2) = 0.428 \text{ and } P = 27.4\cdot P_0.$$

The 8 atoms at the corners of the cube therefore radiate into the y-direction 27.4 times the radiation power of a single atom!

Question: Why does this not violate the energy conservation law?

With a total scattering cross section

$$\sigma_{tot} = 10^{-30}\,m^2 = \sigma_0\cdot\int_\Omega \sin^2\upsilon\,d\Omega$$
$$= \sigma_0\int_{\upsilon=0}^{\pi}\int_{\varphi=0}^{2\pi}\sin^2\upsilon\,d\upsilon\,d\varphi = \pi^2$$

it follows: $\sigma_0 = \sigma_{tot}/\pi^2 \approx 10^{-31}\ m^2$. ($\sigma_0$ is the cross section for the scattering into the solid angle $d\Omega = 1$ sterad around the y-direction ($\vartheta = 90^\circ$). The scattered power for the incident intensity I_i is

$$\Rightarrow P_0(\upsilon = 90^\circ)d\Omega = I_e\cdot\sigma_0 d\Omega$$
$$= 10^{-35}\,m^2\cdot I_e d\Omega.$$

8.4 If the E-vector of the incident wave lies in the plane of incidence we get for the parallel component E the continuity condition at the interface between the two media

$$A_{e\|}\cos\alpha - A_{r\|}\cos\alpha = A_{g\|}\cos\beta.$$

The continuity of the parallel components of the magnetic field $B_\|$ gives the condition (analog to (8.59b)) for non-ferromagnetic media ($\mu_1 \approx \mu_2 \approx 1$)

$$\frac{1}{c'}A_{e\|} + \frac{1}{c_1'}A_{r\|} = \frac{1}{c_2'}A_{g\|}$$
$$\Rightarrow A_{e\|}\cos\alpha - A_{r\|}\cos\alpha = \frac{c_2'}{c_1'}\cos\beta A_{e\|} + \frac{c_2'}{c_1'}\cos\beta A_{r\|}$$
$$\Rightarrow \frac{A_{r\|}}{A_{e\|}} = \frac{\cos\alpha - \frac{c_2'}{c_1'}\cos\beta}{\cos\alpha = \frac{c_2'}{c_1'}\cos\beta}$$

With $c_1'/c_2' = n_1/n_2$ we get

$$\frac{A_{r\parallel}}{A_{e\parallel}} = \frac{n_2 \cos\alpha - n_1 \cos\beta}{n_2 \cos\alpha - n_1 \cos\beta}.$$

8.5 The Fresnel formulas for the amplitude coefficients of the reflected wave at the boundary to a medium with complex refractive index (Fig. A.30) are for the polarization parallel and vertical to the incidence plane

$$\varrho_\perp = \frac{\cos\alpha - (n_2' - \mathrm{i}\,\kappa)\cos\beta}{\cos\alpha + (n_2' - \mathrm{i}\,\kappa)\cos\beta},$$
$$\varrho_\parallel = \frac{(n_2' - \mathrm{i}\,\kappa)\cos\alpha - \cos\beta}{\mathrm{c}(n_2' - \mathrm{i}\,\kappa)os\alpha + \cos\beta}.$$

With $\alpha = 0° \Longrightarrow \varrho_\perp = \varrho_\parallel = \varrho\ , \beta = 0°$

$$\varrho = \frac{1 - (n_2' - \mathrm{i}\,\kappa)}{1 + (n_2' - \mathrm{i}\,\kappa)} = \frac{1 - \kappa^2 + \mathrm{i}\cdot 2\kappa}{(1 + n_2')^2 + \kappa^2}.$$

Numerical example: $\kappa = 2.94$; $n_2' = 0.17$

$$\varrho = \frac{1 - 8.64 + 5.88\cdot\mathrm{i}}{10} = -0.76 + 0.59\,\mathrm{i},$$

$$R = \varrho\cdot\varrho^* = 0.76^2 + 0.59^2 = 0.926.$$

For oblique incidence ($\alpha \neq 0$) we have to determine the refraction angle β in order to calculate the refraction coefficients. We have to extend Snellius' refraction law (8.58) to the refraction at the boundary air-absorbing medium.

The tangential component k_x of the wave vector k stays constant at the boundary from air ($n = 1$) to medium 2 ($n_2 = n' - i\cdot\kappa$). It is

$$\boldsymbol{k}_g = \{k_{gx}, k_{gy}, 0\},$$
$$\boldsymbol{k}_g = \left(\frac{\omega}{c}\right)\{n_1 \sin\alpha, n_2 \cos\beta, 0\}$$

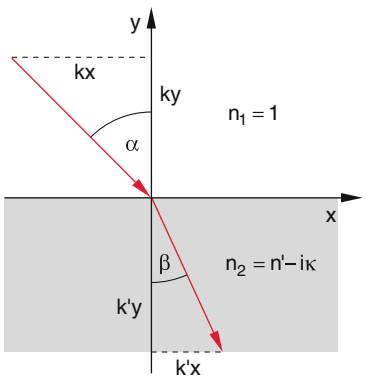


Fig. A.30 Illustration to solution 8.5

with $n_1 = 1$ and $n_2 = n' - i\cdot\kappa$. With the relation

$$\cos\beta = \sqrt{1 - \sin^2\beta}$$

and

$$\sin\beta = \frac{n_1}{n_2}\sin\alpha$$

we get

$$n_2\cos\beta = \sqrt{n_2^2 - \sin^2\alpha} = \eta\cdot\mathrm{e}^{-\mathrm{i}\gamma},$$

where we have expressed the complex quantity $n_2\cdot\cos\beta$ as $\eta\cdot\mathrm{e}^{-\mathrm{i}\gamma} = \eta\cdot(\cos\gamma - \mathrm{i}\cdot\sin\gamma)$. With the comparison of real- and imaginary part we get after squaring

$$\begin{aligned} &n'^2 - \kappa - \sin^2\alpha = \eta^2\cos 2\gamma, \\ &2n'\kappa = \eta^2\sin^2\gamma \\ &\Rightarrow k_g = \frac{\omega}{c}\{\sin\alpha, (\eta\cos\gamma - \mathrm{i}\,\eta\sin\gamma)\}. \end{aligned} \tag{1}$$

For the wave penetrating into the absorbing medium we obtain:

$$\mathrm{e}^{-\mathrm{i}\boldsymbol{k}_g\cdot\boldsymbol{r}} = \mathrm{e}^{[-\mathrm{i}(\omega/c)(\sin\alpha\cdot x) + \eta\cos\gamma\cdot y]}\cdot\underbrace{\mathrm{e}^{[-(\omega/c)\eta\sin\gamma\cdot y]}}_{\text{Absorption}}$$
$$= \mathrm{e}^{-(\alpha/2)y}\cdot\mathrm{e}^{\mathrm{i}(ax+by)}.$$

The surfaces of constant amplitude with y = const. are the surfaces parallel to the interface, the surfaces of constant phase are given by $\sin\alpha\cdot x + \eta\cdot\cos\gamma\cdot y =$ const. They depend on the angle α of incidence and for $\alpha \neq 0$ they do not coincide with the surfaces of equal amplitude. The normal to the phase front point into the direction of the vector

$$\boldsymbol{n}_T = \sin\alpha\cdot\hat{\boldsymbol{x}} + \eta\cos\gamma\cdot\hat{\boldsymbol{y}}$$

They have the amount

$$\sqrt{\sin^2\alpha + \eta^2\cos^2\gamma} = n_T.$$

We define the refraction angle β_T by

$$\sin\beta_T = \frac{\sin\alpha}{\sqrt{\sin^2\alpha + \eta^2\cos^2\gamma}}$$

Now we can write Snellius' refraction law as

$$\frac{\sin\alpha}{\sin\beta_T} = \frac{n_T}{n_1 =}n_T$$

because $n_1 = 1$. Instead of the angle β for transparent media we have to use the angle β_T for absorbing media.

Numerical example:

For the previous example with $n_2' = 0.17$; $\kappa_2 = 2.94$ we obtain from (1)

$$\eta^2 = \sqrt{\left(n_2'^2 - \kappa_2 - \sin^2\alpha\right)^2 + 4n'^2\kappa^2}$$
$$\Rightarrow \eta^2 = 2.42$$
$$\Rightarrow \eta = 1.556,$$
$$\sin 2\gamma = \frac{2n'\kappa}{\eta^2} = 0.413$$
$$\Rightarrow \gamma = 12.2^\circ \Rightarrow \cos^2\gamma = 0.955.$$

For $\alpha = 45^\circ$ this gives

$$\sin\beta_T = \frac{0.71}{\sqrt{0.71^2 + 2.42 \cdot 0.955}}$$
$$= 0.46$$
$$\Rightarrow \beta_T = 27.7^\circ.$$

Replacing in the Fresnel formulas $\cos\beta \rightarrow \cos\beta_T = 0.885$ we obtain

$$\varrho_\perp = \frac{\cos 45^\circ - \left(n_2' - i\kappa\right)\cos\beta_T}{\cos 45^\circ + \left(n_2' - i\kappa\right)\cos\beta_T}$$
$$= \frac{0.71 - 0.17 \cdot 0.885 + i \cdot 2.94 \cdot 0.885}{0.71 + 0.17 \cdot 0.885 - i \cdot 2.94 \cdot 0.885}$$
$$= \frac{0.56 + i \cdot 2.6}{0.86 - i \cdot 2.6}$$
$$\Rightarrow R_\perp = \frac{0.56^2 + 2.6^2}{0.86 + 2.6^2} = \frac{7.07}{7.5}$$
$$\Rightarrow R_\perp = 0.943.$$

The corresponding calculations yield $\varrho_\parallel$ and $R_\parallel$ for $\alpha = 45^\circ$ and $\alpha = 85^\circ$.

8.6 $P(x) = P_0 \cdot e^{-\alpha x}$.
The absorbed power is

$$\Delta P = P_0 - P(x) = P_0(1 - e^{-\alpha x}).$$

For $\alpha = 10^{-3}$ cm^{-1}, $x = d = 3$ cm is

$$\Delta P \approx P_0 \cdot \alpha d = 3 \times 10^{-3} P_0.$$

Only 0.3% are absorbed.
The situation changes for $\alpha = 1$ cm^{-1} and $d = 3$ cm

$$\Delta P = P_0\left(1 - e^3 -\right) = 0.95 P_0.$$

8.7

$$\sin\alpha = \frac{R - d/2}{R + d/2} \geq \sin\alpha_g = \frac{n_2}{n_1}$$
$$\Rightarrow R - \frac{d}{2} \geq \frac{n_2}{n_1}\left(r + \frac{d}{2}\right)$$
$$\Rightarrow R \geq \frac{d}{2}\frac{1 + n_2/n_1}{1 - n_2/n_1} = \frac{d}{2}\frac{n_2 + n_1}{n_2 - n_1}.$$

For $d = 10$ µm, $n_1 = 1.6$; $n_2 = 1.59 \Longrightarrow$
$R \geq 5 \times 10^{-6} \cdot \frac{3.1}{0.01}$ m $= 1550$ µm $= 1.5$ mm.

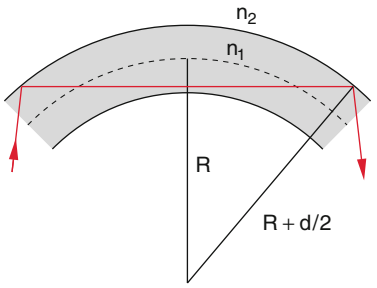


Fig. A.31 Illustration to solution 8.7

8.8 For $\omega - \omega_0 \gg \gamma$ we get from (16) with

$$a_1 = \frac{Ne^2}{2\varepsilon_0 m}, \quad a_2 = \frac{a_1}{4\pi^2 c^2},$$
$$n - 1 \approx \frac{a_1}{\omega_0^2 - \omega^2} = \frac{a_2}{1/\lambda_0^2 - 1/\lambda^2}$$
$$= \frac{a_2\lambda_0^2\lambda^2}{\lambda^2 - \lambda_0^2} = a_2\lambda_0^2 + \frac{a_2\lambda_0^4}{a_2\lambda_0^2}$$
$$= a + \frac{b}{\lambda^2 - \lambda_0^2},$$

with $a = a_2 \cdot \lambda_0^2$ and $b = a_2 \cdot \lambda_0^4$.

8.9 (a) The extra-ordinary refractive index $n_a(\theta)$ obeys the equation for an ellipse

$$\frac{1}{n_e^2(\theta)} = \frac{\cos^2\theta}{n_0^2} + \frac{\sin^2\theta}{n_a^2(\theta = 90^\circ)}. \tag{1}$$

Phase matching is achieved for $n_0(\omega) = n_e(2\omega)$

$$\Rightarrow \frac{1}{n_a^2(\theta, 2\omega)} = \frac{1}{n_0^2\omega} = \frac{1 - \sin^2}{n_0^2(2\omega)} + \frac{\sin^2\theta}{n_a^2(2\omega)}$$
$$\Rightarrow \sin^2\theta_{opt} = \frac{[n_0\omega]^{-2} - [n_0(2\omega)]^{-2}}{[n_a(2\omega)]^{-2} - [n_0(2\omega)]^{-2}}. \tag{2}$$

Inserting the numerical values yields

$$\sin^2\theta_{opt} = 0.5424 \Rightarrow \theta_{opt} = 47.4^\circ.$$

(b) For $\theta = 48.8^\circ$ is according to (1) $n_e(48.80.\ 2\omega) = 1.674$. The difference $\Delta n = n_0(\omega) - n_e(\theta,\ 2\omega)$

decreases to 0.001 and the coherence length becomes

$$L_{\text{kohärent}} = \frac{\lambda/2}{|n_a(2\omega) - n_0(\omega)|} = 500\lambda = 250\,\mu\text{m}.$$

(c) The output intensity $I(2\omega, L)$ becomes for $\Delta h = (2\pi/\lambda)\Delta n = 1.25 \times 10^4\ \text{m}^{-1}$

$$I(2\omega, L) = \frac{10^{24} \cdot 2 \cdot 3.5^2 \times 10^{30} \cdot 64 \times 10^{-24} \cdot 2.5^2 \times 10^{-8}}{1.675^3 \cdot 27 \times 10^{24} \cdot 8.85 \times 10^{-12}} = 1.5 \times 10^{11}\text{W/m}^2.$$

This is about 15% of the input intensity I_i of the fundamental wave.

Chapter 9

9.1 We will show that an incident plane wave propagating into the x-direction, is focused by a parabolic mirror into the point F. This can be proved by showing that for all rays of the wave the optical path lengths from the plane $x = f$ until the point $F = \{f, 0\}$ are equal independent of y.

$$\begin{aligned} s &= s_1 + s_2 \\ &= (f - x) + \sqrt{y^2 + (f - x)^2} = \min \\ \Rightarrow \frac{ds}{dx} &= -1 + \frac{2yy' - 2(f - x)}{2 \cdot \sqrt{y^2 + (f - x)^2}} = 0 \\ \Rightarrow yy' - (f - x) &= \sqrt{y^2 + (f - x)^2} \\ y' - \frac{f - x}{y} &= \sqrt{1 + \left(\frac{f - x}{y}\right)^2} \end{aligned}$$

Squaring gives

$$y'^2 - \frac{2(f - x)}{y} y' = 1.$$

The solution of this equation is ω

$$\Rightarrow y = \sqrt{4fx} \Rightarrow y^2 = 4fx \Rightarrow 2yy' = 4f.$$

9.2 (a) If a plane mirror is turned by the angle δ the angle of incidence changes from α to $\alpha + \delta$ and the reflection angle also changes from α to $(\alpha + \delta)$, The deflection angle of the reflected beam against the incident beam is therefore $\Delta = 2\alpha + 2\delta$, which is larger by 2δ against the mirror before its turning (Fig. A.32a).

(b) For the spherical mirror, however, there is no change of the direction of the reflected beam, when the mirror is turned around the center of curvature (Fig. A.32b). If, however, the spherical mirror is turned around the point where the light is incident onto the mirror, the direction of the reflected light beam is also turned by the angle 2δ exactly as for the plane mirror.

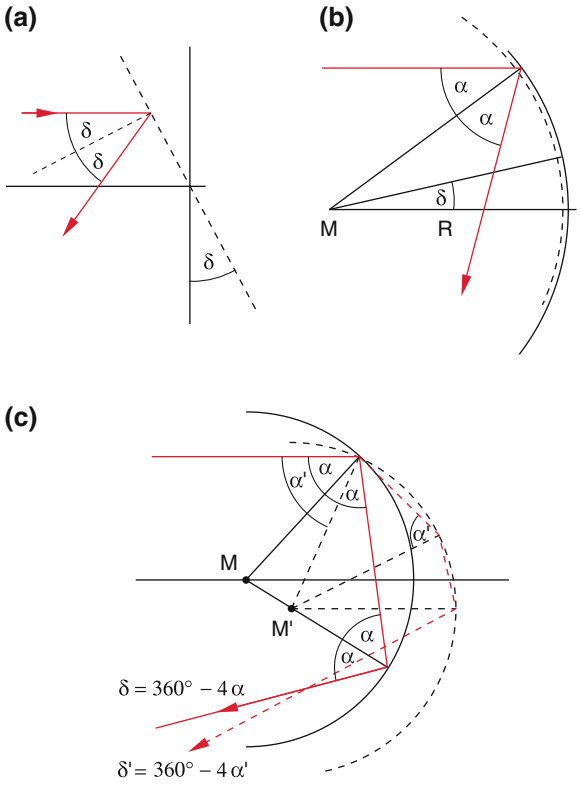


Fig. A.32 Illustration to solution 9.2

For a twofold reflection the total deflection is $\Delta = 360° - 4\alpha$ for the incidence angle α and $\Delta' = 360° - 4(\alpha+\delta)$ if the mirror is turned by the angle δ. Furthermore a beam shift occurs (Fig. A.32c).

9.3 From Fig. 9.27 we can find

$$\begin{aligned} \tan\alpha &= \frac{G}{a} = \frac{B}{b} \Rightarrow G = \frac{a}{b} \cdot B, \\ \tan\beta &= \frac{G}{f} = \frac{B}{b - f} \Rightarrow \frac{a \cdot B}{b \cdot f} = \frac{B}{b - f} \\ &\Rightarrow ab = af = bf, \\ f &= \frac{ab}{a + b} \Rightarrow \frac{1}{f} = \frac{1}{a} + \frac{1}{b}. \end{aligned}$$

9.4 The virtual images B_i of the point A which are generated by the reflection at the mirrors M_1 and M_2 are located at the positions x_i. We get

$$B_1: x_1 = -\frac{d}{2} - \frac{d}{3} = -\frac{5}{6}d,$$
$$B_2: x_2 = \frac{d}{2} + -\frac{2}{3}d = \frac{7}{6}d,$$
$$B_3: x_3 = \frac{d}{2} + \frac{d}{2} + \frac{d}{6}d = \frac{11}{6}d,$$
$$B_4: x_4 = -\frac{13}{6}d.$$

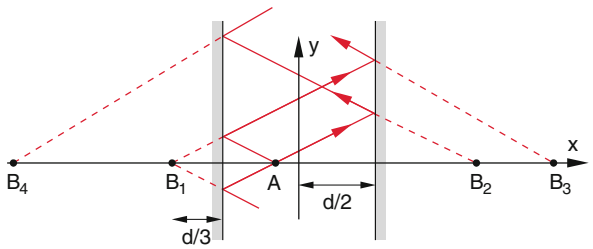


Fig. A.33 Illustration to solution 9.4

9.5 From the relations

$$\frac{\sin\alpha}{\sin\beta} = n_2; \quad \frac{\sin\gamma}{\sin\beta} = \frac{n_2}{n_1}$$
$$\Rightarrow \sin\gamma = \frac{n_2}{n_1}\sin\beta = \frac{1}{n_1}\sin\alpha,$$
$$n_1 = 1.46, \quad n_2 = 1.33,$$
$$h_1 = 4\,\text{cm}, \quad h_2 = 2\,\text{cm}.$$

we get

(a) $\alpha_m = 90°$: at the upper interface total reflection occurs.

$$\Rightarrow \sin\beta_m = \frac{1}{n_2} = 0.752 \Rightarrow \beta_m = 48.76°$$
$$\Rightarrow \sin\gamma_m = \frac{1}{n_1} = 0.685 \Rightarrow \gamma_m = 43.235°.$$

The radius R of the cylindrical glass container has to be

$$R \geq x_1 + x_2 = h_1 \cdot \tan\gamma_m + h_2 \cdot \tan\beta_m$$
$$4\,\text{cm} \cdot \tan 43.23° + 2\,\text{cm} \cdot \tan 48.76°$$
$$6.04\,\text{cm}.$$

(b) For $R \leq 6.04$ cm the maximum observable angle can be obtained from

$$R = x_1 + x_2 = h_1 \tan\gamma + h_2 \tan\beta$$
$$h_1 \frac{\sin\gamma}{\cos\beta} + h_2 \frac{\sin\beta}{\cos\beta}$$
$$= \frac{h_1}{n_1} \frac{\sin\alpha}{\sqrt{1 - 1/n_1^2 \cdot \sin\alpha}}$$
$$+ \frac{h_2}{h_1} \frac{\sin\alpha}{\sqrt{1 - 1/n_1^2 \cdot \sin\alpha}}$$
$$= \frac{h_1 \cdot \sin\alpha}{\sqrt{1 - n_2^2/n_1^2 \cdot \sin\alpha}} + \frac{h_2 \cdot \sin\alpha}{\sqrt{1 - \sin^2\alpha}}.$$

More elegant is the solution based on Fermat's principle. For the light travel time T is

$$T^2 = \frac{x_1^2 + h_1^2}{n_1^2 \cdot c^2} + \frac{x_2^2 + h_2^2}{n_2^2 \cdot c^2} = \min.$$

With $x_2 = R - x_1$ it follows (Figs. A.33 and A.34):

$$\frac{dT^2}{dx_1} = \frac{2x_1}{c \cdot n_1^2} - \frac{2(R - x_1)}{c \cdot n_2^2} = 0$$
$$x_1 = \frac{n_1^2}{n_2^2} \cdot (R - x_1)$$
$$\Rightarrow x_1 = R \cdot \frac{1}{1 + n_2^2/n_1^2} = \frac{R}{1.83}$$
$$\Rightarrow \tan\gamma = \frac{x_1}{h_1} = \frac{R}{1.83 h_1} = 0.41$$
$$\Rightarrow \gamma = 22.3° \Rightarrow \sin\gamma = 0.38$$
$$\Rightarrow \sin\beta = \frac{n_1}{n_2}\sin\gamma = 0.417$$
$$\Rightarrow \beta = 24.6°$$
$$\Rightarrow \alpha = 33.6°.$$

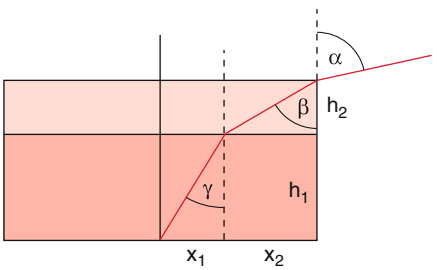


Fig. A.34 Illustration to solution 9.5

9.6 With the lens equation

$$\frac{1}{a} + \frac{1}{b} = \frac{1}{f}$$

and the image scale $B/A = b/a = 10$ and $a + b = 3$ m it follows

$$11a = 3\,\text{m} \Rightarrow a = \frac{3}{11}\,\text{m},$$
$$b = \left(3 - \frac{3}{11}\right)\text{m} = \frac{30}{11}\,\text{m}$$
$$\Rightarrow f = \frac{a \cdot b}{a + b} = \frac{90}{11 \cdot 11 \cdot 3}\,\text{m} = 0.25\,m.$$

9.7 The incident beam is deflected downwards at the first interface by the angle $(\alpha - \beta)$, at the second interface by the angle $-(\alpha - ß)$ upwards. The total deflection is then$\varphi = (\alpha - \beta) - (\alpha - \beta) = 0$. This means that the exit beam is parallel to the incident beam, but it is shifted (Fig. A.35). The shift is

$$\Delta = \frac{d}{\cos\beta} \cdot \sin(\alpha - \beta)$$
$$\frac{d}{\sqrt{1-\sin^2\beta}}(\sin\alpha\cos\beta - \cos\alpha - \sin\beta)$$
$$= \frac{d \cdot n}{\sqrt{n^2 - \sin^2 a}}$$
$$\sin\alpha\left(\sqrt{1 - \frac{\sin^2\alpha}{n^2}} - \frac{1}{n}\cos\alpha\,\sin\alpha\right)$$
$$= \frac{d \cdot \sin\alpha}{\sqrt{n^2 - \sin^2\alpha}}\left(\sqrt{n^2 - \sin^2\alpha} - \cos\alpha\right)$$
$$= d \cdot \sin\alpha\left(1 - \frac{\cos\alpha}{\sqrt{n^2 - \sin^2\alpha}}\right).$$

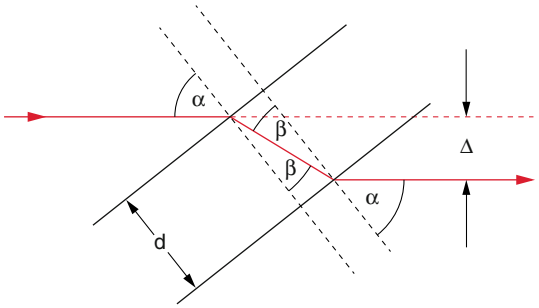


Fig. A.35 Illustration to solution 9.7

9.8 At first we regard a light beam in the x-y-plane, which impinges onto the mirror under the angle α. Its total deflection $\Delta\varphi$ is with $\beta = 90° - \alpha$

$$\Delta\varphi = 2\beta + 2\alpha = 2(90° - \alpha) + 2\alpha = 180°.$$

If the beam propagates inclined against the x-y-plane we partition the wave vector $\mathbf{k}$ into a parallel component $k_{||} = \{k_x, k_y, 0\}$ and $k_{||} = \{0, 0, k_z\}$. For $k_{||}$ the consideration above is valid. Since the two mirrors in the x-y plane are perpendicular to each other the component k_z is transferred into $-k_z$ after two reflections. This means that for any incidence angle α the wave vector $\mathbf{k}$ is transferred into $-\mathbf{k}$, which means that the incident beam is back-reflected into the incidence direction (Fig. A.36).

9.9 Generally the distance between the two lenses of a telescope is chosen as $d = f_1 + f_2$, which causes parallel light reaching the eye of the observer.
According to the intercept theorem is

$$D_1/D_2 = f_1/f_2.$$

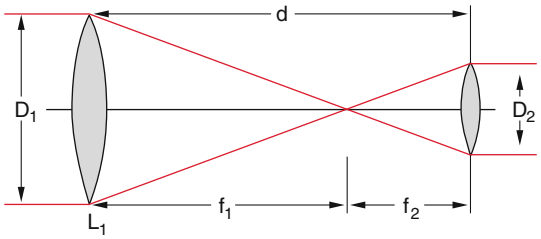


Fig. A.36 Illustration to solution 9.9

The diameter of the ocular lens has to be

$$D_2 = D_1 \cdot \frac{f_2}{f_1} = 5 \cdot \frac{2}{20}\text{cm} = 0.5\,\text{cm}$$

in order to transmit the whole incident beam. The angular magnification of the telescope is

$$V = \frac{f_1}{f_2} = 10$$

(see Sect. 11.2.3).

9.10 (a) For the triangle MAP is the sine theorem

$$\frac{x_2}{R} = \frac{\sin\beta}{\sin(90° + \beta + \gamma)} = \frac{\sin\beta}{\sin(\alpha - \beta)}.$$

An intersection point exists only for $x_2 < R$.

$$\Rightarrow \sin\beta < \sin(\alpha - \beta)$$
$$\Rightarrow \frac{\sin\beta}{n} < \sin(\alpha - \beta)$$
$$\Rightarrow \frac{h}{R} < n \cdot \sin(\alpha - \beta)$$
$$\Rightarrow h < R \cdot n \cdot \sin(\alpha - \beta)$$

Using the relation

$$\sin(\alpha - \beta) = \sin\alpha\cos\beta - \cos\alpha\,\sin\beta$$
$$\frac{h}{R}\sqrt{1 - \frac{\sin^2\alpha}{n^2}} - \frac{\cos\alpha\,\sin\alpha}{n}$$

this can be written as

$$h < R \cdot \sqrt{n^2 - (1 + \cos\alpha)^2}.$$

(b) As can be seen from Fig. A.37a the total deflection angle is

$$\delta = \alpha - \beta + (360° - 2\beta) + \alpha - \beta$$
$$= 360° + 2\alpha - 4\beta.$$

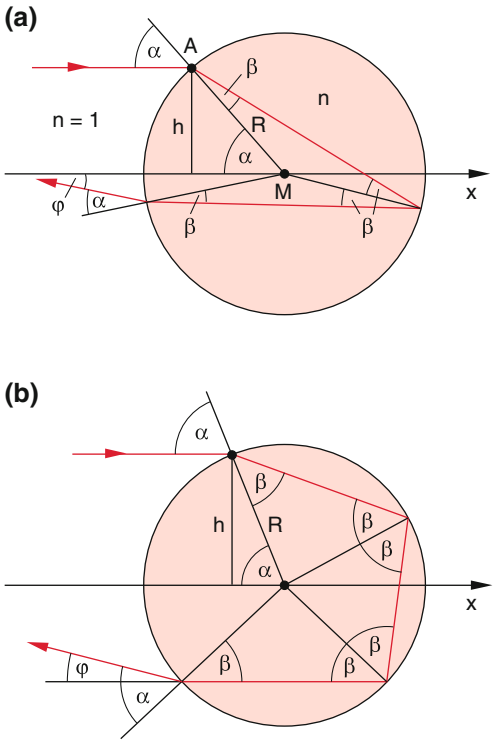


Fig. A.37 Illustration to solution 9.10

Measured against the backward direction the deflection is

$$\varphi = \delta - 180^\circ = 180^\circ + 2\alpha - 4\beta.$$

With $\sin\alpha = h/R$ and $\sin\beta = (h/R)/n$ we get

$$\varphi - 180^\circ + 2\arcsin\frac{h}{R} - 4\arcsin\left(\frac{1}{n}\cdot\frac{h}{R}\right).$$

(c) The deflection angle becomes minimum for $d\varphi/dh = 0$.

$$\frac{d\varphi}{dh} = \frac{2\sqrt{R}}{\sqrt{1-h^2/R^2}} - \frac{4/(n\cdot R)}{\sqrt{1-h^2/(n^2R^2)}} = 0$$

$$h_m = R\cdot\sqrt{\frac{1}{3}(4-n^2)}$$

$$\Rightarrow \sin\alpha_m = \frac{h}{R} = \sqrt{\frac{1}{3}(4-n^2)}$$

(d) With $n = 1.33$ we get the numerical results

$$\sin\alpha_m = 0.86238 \Rightarrow \alpha_m = 59.6^\circ,$$

$$\sin\beta_m = \frac{\sin\alpha_m}{n} = 0.6484 \Rightarrow \beta_m = 40.4^\circ$$

$$\Rightarrow \varphi = 180^\circ + 2\alpha - 4\beta = 137.6^\circ.$$

After the second reflection the total deflection is

$$\delta = 360^\circ + 2\alpha - 6\beta,$$

This means t at the deflection against the backwards direction is

$$\varphi = 180^\circ + 2\alpha - 6\beta$$
$$= 180^\circ + 2\arcsin(h/R) - 6\arcsin\left(\frac{1}{n}\cdot\frac{h}{R}\right).$$

For the minimum deflection is $d\varphi/dh = 0$. This gives the relations

$$h_m = R\cdot\sqrt{\frac{1}{8}(9-n^2)}$$
$$\Rightarrow \frac{h_m}{R} = 0.951 \Rightarrow \varphi_m = 128^\circ.$$

9.11 (a) According to (9.25a) we obtain

$$f = \frac{1}{n-1}\frac{R_1R_2}{R_2-R_1}$$
$$n(600\,\text{nm}) = 1.485$$

The refractive index for $\lambda = 600$ nm is n (600 nm) = 1.48

$$f_{\text{rot}} = \frac{1}{0.485}\cdot\frac{200}{10}\,\text{cm} = 41.24\,\text{cm},$$

With the refractive index n (400 nm) = 1.50 the focal length becomes

$$\text{f}_{\text{blau}} = \frac{1}{0.50}\cdot 20\,\text{cm} = 40\,\text{cm}.$$

(b) In order to compensate the chromatic aberration a diverging lens with focal length f_2 has to be used, which can be calculated as follows: According to (9.34d) the ratio of the focal lengths $f_2(n_g)/f_1(n_g)$ with

$$n_g = \frac{1}{2}(n_r + n_b) = 1.492$$

must be

$$\frac{f_2}{f_1} = -\frac{(n_{1\text{g}}-1)(n_{2\text{b}}-n_{2\text{r}})}{(n_{2\text{g}}-1)(n_{1\text{b}}-n_{1\text{r}})}$$
$$= -\frac{0.492\cdot(n_{2\text{b}}-n_{2\text{r}})}{(n_{2\text{g}}-1)\cdot(1.5-1.485)}$$
$$= -\frac{32.8\cdot(n_{2\text{b}}-n_{2\text{r}})}{1/2(n_{2\text{b}}-n_{2\text{r}})-1}.$$

Choosing $n_{2\text{b}} = 1.6$ and $n_{2\text{r}} = 1.55$ we obtain

$$\frac{f_2}{f_1} = -2.85.$$

The focal length of the diverging lens in the achromat must therefore be

$$f_2 = -2.85 f_1 = -2.85 \cdot 40.62\,\text{cm} = -115.85\,\text{cm}.$$

9.12 Since the distance D between the two lenses is smaller than their focal lengths f_1 and f_2 the focal length of the total system is obtained, according to (9.32) from the relation

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{D}{f_1 f_2} = \left(\frac{1}{10} + \frac{1}{50} + \frac{5}{500}\right)\frac{1}{\text{cm}} = \frac{55}{500}\frac{1}{\text{cm}}$$
$$\Rightarrow f = 9.1\,\text{cm}.$$

9.13 We start from the imaging equation

$$\frac{1}{g} + \frac{1}{b} \approx \frac{2}{R}.$$

For the imaging by the spherical mirror S_1 we get (Fig. A.38a)

$$g_1 = \overline{S_1 A} = x = 6\,\text{cm}, \quad \text{R}_1 = 24\,\text{cm}$$
$$\Rightarrow b_1 = \frac{g_1 R_1}{2g_1 - R_1} = \frac{2 \cdot 6 \cdot 24}{12 - 24}\,\text{cm} = -24\,\text{cm}.$$

The imaging is divergent because A is located between mirror S_1 and focal point F_1. A virtual image B^* is generated on the left side of S_1 at the distance $x = -24$ cm from S_1.

For the imaging by S_2 we get

$$g_2 = -(d - x) = 54\,\text{cm},$$
$$\text{R}_2 = -40\,\text{cm}$$
$$\Rightarrow b_2 = \frac{54 \cdot 40\,\text{cm}}{-2 \cdot 54 + 40} = -31\,\text{cm}$$
$$\Rightarrow x(b_2) = (60 - 31)\,\text{cm} = 29\,\text{cm}.$$

B_2 can be again imaged by S_1 into B_3. We get

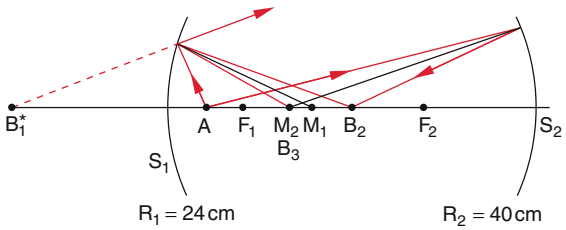


Fig. A.38 Illustration to solution 9.13

$$b_3 = \frac{g_3 R_1}{2g_3 - R_1} \text{ with } \text{g}_3 = 29\text{cm}$$
$$\Rightarrow b_3 = 20\text{cm}.$$

This is identical with the center M_2 of the spherical mirror S_2. Therefore B_3 is imaged by S_2 into itself, by S_1 into B_2, etc. There are therefore two real images and one virtual image.

9.14 The matrix of the system is

$$\widetilde{M} = \widetilde{B}_7 \cdot \widetilde{T}_{76} \cdot \widetilde{B}_6 \cdot \widetilde{T}_{65} \cdot \widetilde{B}_5 \cdot \widetilde{T}_{54} \cdot \widetilde{T}_4 \cdot \widetilde{T}_{43} \cdot \widetilde{B}_3 \cdot \widetilde{T}_{32} \cdot \widetilde{B}_2 \cdot \widetilde{T}_{21} \cdot \widetilde{B}_1.$$

$$\widetilde{B}_1 = \begin{pmatrix} 1 & -\dfrac{1.6116 - 1}{1.628} \\ 0 & 1 \end{pmatrix}, \quad \widetilde{B}_2 = \begin{pmatrix} 1 & -\dfrac{1 - 1.6116}{-27.57} \\ 0 & 1 \end{pmatrix}$$

The translation matrices are

$$\widetilde{T}_{21} = \begin{pmatrix} 1 & 0 \\ \dfrac{0.357}{1.6116} & 1 \end{pmatrix}, \quad \widetilde{T}_{32} = \begin{pmatrix} 1 & 0 \\ \frac{0.189}{1} & 1 \end{pmatrix}$$

etc. The product matrix, which can be best calculated with a computer program, is

$$\widetilde{M} = \begin{pmatrix} 0.848 & -0.198 \\ 1.338 & 0.867 \end{pmatrix}.$$

With the approximation of thin lenses we get according to (9.45a) $M_{12} = -1/\text{f}$ which gives $f = 5.06$ cm.

Chapter 10

10.1 We start from (10.5)

$$\Delta s + \sqrt{(x - d)^2 + y^2 + z_0^2} = \sqrt{(x - d)^2 + y^2 + z_0^2}.$$

Squaring and cancelling gives

$$4xd - \Delta s^2 = 2\Delta s \sqrt{(x - d)^2 + y^2 + z_0^2}.$$

If we again square this equation and rearrange the terms we get

$$x^2(16d^2 - 4\Delta s^2) = 4\Delta s^2(d^2 + y^2 + z_0^2 - \Delta s^2)$$
$$\Rightarrow \frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

With

$$a^2 = \frac{d^2 + z_0^2 - \Delta s^2}{(2d/\Delta s^2) - 1},$$
$$b^2 = d^2 + z_0^2 - \Delta s^2.$$

The distance between the vertices of the two hyperbolas is

$$\Delta x_s = 2a.$$

For $z_0 \gg d$ we get

$$\Delta s = z_0 \left[\sqrt{1 + \frac{(x+d)^2}{z_0^2} + \frac{y^2}{z_0^2}} - \sqrt{1 + \frac{(x+d)^2}{z_0^2} + \frac{y^2}{z_0^2}} \right]$$
$$\Rightarrow \Delta s \approx z_0 \left(\frac{2xd}{z_0^2} \right) = \frac{2xd}{z_0} = m \cdot \lambda.$$

and we obtain the distances between the two vertices

$$\Delta x_S = 2a = \frac{z_0}{d} \cdot m \cdot \lambda.$$

10.2 The optical path difference between the two arms of the Michelson interferometer is

$$\Delta s = \Delta s_1 - \Delta s_2$$

with

$$\Delta s_1 = \frac{d_1}{\cos\alpha} + \frac{\Delta x}{\cos\alpha},$$

and

$$\Delta x = d_1 - (y_1 + y_2),$$
$$y_1 = d_1 \tan\alpha, \quad y_2 = d_1 - (y_1 + y_2)\tan\alpha$$
$$\Rightarrow y_2 = d_1 \cdot \frac{\tan\alpha(1 - \tan\alpha)}{1 + \tan\alpha}$$
$$\Rightarrow \Delta x = d_1 \cdot \frac{1 - \tan\alpha)}{1 + \tan\alpha}$$
$$\Rightarrow \Delta s_1 = \frac{2d_1}{\cos\alpha} \frac{1}{1 + \tan\alpha} = \frac{2d_1}{\cos\alpha + \sin\alpha}.$$

In the same way we get

$$\Delta s_2 = \frac{2d_2}{\cos\alpha} \frac{1}{1 + \tan\alpha} = \frac{2d_2}{\cos\alpha + \sin\alpha}.$$

Note, that the beam splitter in inclined by 45°. This implies $\Delta x = \Delta y$ for $d_1 = d_2$ (Fig. A.39).

The path difference between the two partial beams which are inclined by the angle α against the symmetry axis is then

$$\Delta s = 2\frac{d_1 - d_2}{\cos\alpha + \sin\alpha}.$$

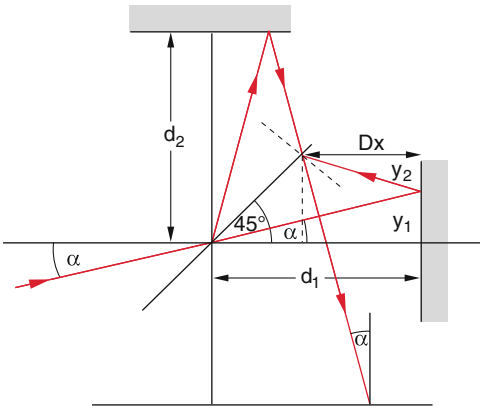


Fig. A.39 Illustration to solution 10.2

For $\Delta s = m \cdot \lambda$ bright rings are observed in the observation plane, which change their radius R when the path difference $d_1 - d_2$ varies. For a given value of the integer m we get

$$\cos\alpha + \sin\alpha = \frac{d_1 - d_2}{m \cdot \lambda/2}$$

10.3 The light beam which is reflected at the mirror M_1 inclined by the angle 2δ against the symmetry axis impinges onto the observation plane under the same angle 2δ against the normal of the plane (Fig. A.40). It is still a plane wave.
The phase difference between the wave with $\delta = 0$ and the inclined wave is

$$\phi(x) = 2\pi \cdot \frac{x}{\lambda} \cdot \sin 2\delta.$$

The fringe separation Δx of the interference pattern occurs for $\Delta\phi = 2\pi$. We therefore get

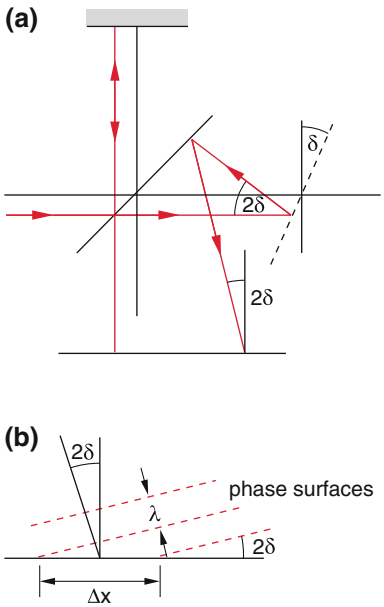


Fig. A.40 Illustration to solution 10.3

$$\Delta x = \frac{\lambda}{\sin 2\delta}.$$

10.4 What is the reflectivity of a dielectric coating for a plane wave which reaches the layer perpendicular ($\alpha = 0$) for a coating with

(a) $n_H d = \lambda/4$,

(b) $n_H d = \lambda/2$,

(c) a coating consisting of two alternate layers with high refractive index n_h and low refractive index n_l and $n_H d = \lambda/4 \sim$ H and $n_L d = \lambda/4 \sim$ L with $n_H = 1.8$ and $n_L = 1.3$ on a substrate with $n_s = 1.5$? Discuss the different effects of λ/4 and λ/2 layers. For which value of n_h is the reflectivity in case (a) completely suppressed?

Solution (from Dr. E. Welsh, University Jena)
Analytical solution for the case of a λ/4-layer (two interfaces AH and HS (Fig. A.41). For vertical incidence there is no polarization dependence.
Approach:

$$E_0 = A_0 e^{ik_0 z} + A_r e^{ik_0 z}, \quad k_0 = \frac{2\pi}{\lambda} \quad \text{in} \tag{1}$$

$$E_H = A_1 e^{ik_H z} + A_2 e^{-ik_H z}, \quad k_H = \frac{2\pi}{\lambda} n_H \quad \text{in} \tag{2}$$

$$E_S = A_t e^{ik_S z} \quad k_S = \frac{2\pi}{\lambda} n_S \quad \text{in} \tag{3}$$

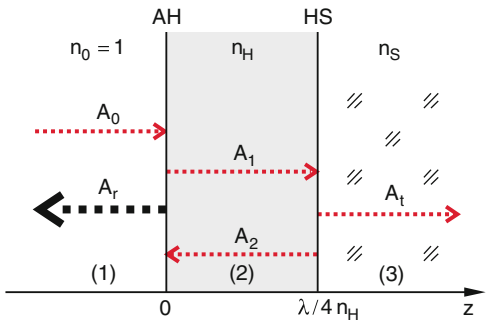


Fig. A.41 Illustration to solution 10.4

For $A_0 = 1$ all coefficients are normalized to the incident intensity. The reflectivity is then $R = A_r^2$. The other 4 unknown amplitudes can be calculated when the continuity of the amplitudes at the two interfaces is observed (Fig. A.42).

The boundary condition at $z = 0$ demands:

$$\begin{aligned} &E_0(z=0) = E_H(z=0) \\ &\Rightarrow 1 + A_r = A_1 + A_2, \end{aligned} \tag{4}$$

$$\begin{aligned} &\frac{d}{dz}E_0(z=0) = \frac{d}{dz}E_n(z=0) \\ &\Rightarrow 1 - A_r = n_H(A_1 - A_2). \end{aligned} \tag{5}$$

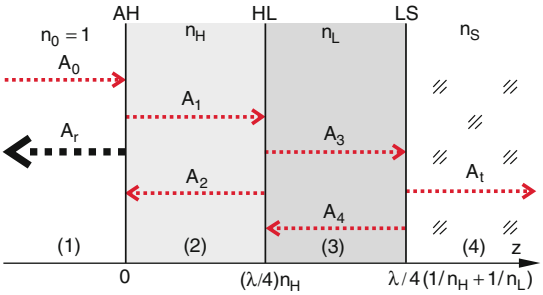


Fig. A.42 Illustration to solution 10.4

The boundary condition at $z = \lambda/(4n_H)$ demands

$$\begin{aligned} &E_H(z = \lambda/4n_H) = E_t(z = \lambda/4n_H) \\ &\Rightarrow iA_1 - iA_2 = A_t e^{i(\pi/2)(n_S/n_H)}, \end{aligned} \tag{6}$$

$$\begin{aligned} &\frac{d}{dz}E_H(z = \lambda/4n_H) = \frac{d}{dz}E_t(z = \lambda/4n_H) \\ &\Rightarrow -n_H(A_1 + A_2) = in_S e^{i(\pi/2)(n_S/n_H)} A_t. \end{aligned} \tag{7}$$

The Eqs. (1)–(4) leads with the abbreviation

$$\delta = e^{i\frac{\pi n_S}{2n_H}}$$

to the system of equations

$$\begin{aligned} -A_r + A_1 + A_2 + 0 &= 1 \\ -A_r + n_H A_1 - n_H A_2 + 0 &= 1 \\ 0 + A_1 - A_2 + i\delta A_t &= 0 \\ 0 + n_H A_1 + n_H A_2 + in_S \delta A_t &= 0 \end{aligned} \tag{8}$$

The coefficient determinant is

$$|D| = i\delta \begin{vmatrix} -1 & 1 & 1 & 0 \\ 1 & n_H & -n_H & 0 \\ 0 & 1 & -1 & 0 \\ 0 & n_H & n_H & n_S \end{vmatrix}. \tag{9}$$

The solution gives for A_r

$$A_r = \frac{|D_R|}{|D|} = \frac{n_S - n_H^2}{n_S + n_H^2} \tag{10}$$

and the reflectivity

$$R = A_r^2 = \left(\frac{n_S - n_H^2}{n_S + n_H^2}\right)^2. \tag{11}$$

Numerical values:

$n_s = 1.5$; $n_h = 1.8$; $\rightarrow A_r^2 = 0.13$:

Discussion:

Because $n_H > n_S$ a phase jump of π occurs for the reflection at the interface at $z = 0$ and the phase difference between the partial waves reflected at $z = 0$ and $z = \lambda/4n_H$ becomes $\Delta\varphi = \pi$. For $n_H < n_S$ an additional phase jump of π at $z = \lambda/4n_H$ occurs. In

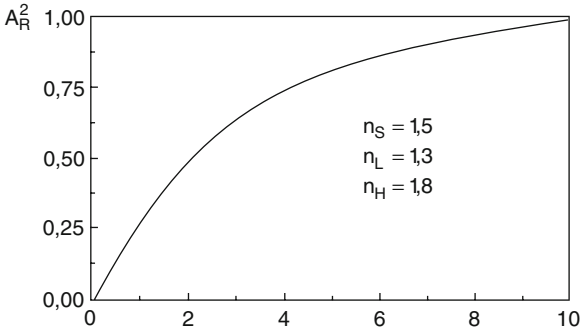


Fig. A.43 Illustration to solution 10.4

this case destructive interference between the two reflected waves at $z = 0$ and $z = \lambda/4n_H$ appears. For $n_H^2 - n_0 n_S = 0$ the reflected amplitude A_r becomes zero. In this case the refractive index of the layer must be $n_H = \sqrt{1.5} \approx 1.22$ instead of 1.8.

(c) The analytical solution for the case of two layers with n_h and n_l (Fig. A.43) has to regard 3 interfaces AH, HL and LS.

The ansatz for the electric field amplitudes is

$$\begin{aligned} E_0 &= A_0 e^{ik_0 z} + A_r e^{-ik_0 z}, \\ E_H &= A_1 e^{ik_H z} + A_2 e^{-ik_H z}, \\ E_L &= A_3 e^{ik_L z} + A_4 e^{-ik_L z}, \\ E_S &= A_z e^{-k_S z}. \end{aligned} \tag{12}$$

The same boundary conditions as in (a) for the three interfaces at $z = 0; z = \lambda/(4n_h)$ and $z = \lambda/(4 \cdot (1/n_h + 1/n_s))$ lead to a 6×6 system of linear equations with a solution that can be obtained analogous to (1). The reflected amplitude is

$$A_r = \frac{n_S - \left(\frac{n_L}{n_L}\right)^2}{n_S - \left(\frac{n_L}{n_H}\right)^2}. \tag{13}$$

and the reflectivity becomes

$$A_r^2 = \left(\frac{n_S - \left(\frac{n_L}{n_L}\right)^2}{n_S - \left(\frac{n_L}{n_H}\right)^2} \right)^2. \tag{14}$$

The numerical value is $A_r^2 \approx 0.23$.

Since the phase jump of π only occurs at the interfaces AH and LS the $\lambda/4$-components $\Delta\varphi = \pi/2$ cause constructive interference. If several HL-layers are laid on top of each other the reflected amplitude can be substantially enhanced. One obtains quantitatively

$$A_r^2 = \left(\frac{n_S - \left(\frac{n_L}{n_H}\right)^{2k}}{n_S - \left(\frac{n_L}{n_H}\right)^{2k}} \right)^2. \tag{15}$$

For $n_L = n_S$ the third interface disappears and the reflectivity becomes

$$A_r = \frac{n_S - 1}{n_S + 1}, \quad A_r^2 = \left(\frac{n_S - 1}{n_S + 1}\right)^2. \tag{16}$$

Numerical value: $R = 0.04$.

10.5 For vertical incidence ($\alpha = 0$) the path difference between two edge rays is for the diffraction angle θ

$$\Delta s = b \cdot \sin\theta.$$

For inclined incidence with $\alpha = \alpha_0$ the path difference becomes

$$\Delta s = b \cdot (\sin\theta - \sin\alpha_0) = \Delta_2 - \Delta_1.$$

Instead of $\sin\theta$ the expression $\sin\theta - \sin\alpha_0$ has to be inserted into Eq. (10.45). The central diffraction maximum appears at $\theta_0 = \alpha_0$.

The ± 1 diffraction maximum appears at the angle θ defined by the equation

$$\begin{aligned} &\frac{b}{\lambda}(\sin\theta - \sin\alpha_0) = \pm 1 \\ &\Rightarrow \sin\theta_{1,2} = \pm\frac{\lambda}{b} + \sin\alpha_0. \end{aligned}$$

The angular width of the central maximum is

$$\begin{aligned} \Delta\theta &= \theta_1 - \theta_2 \\ &= \arcsin\left(\sin\alpha_0 + \frac{\lambda}{b}\right) \\ &\quad - \arcsin\left(\sin\alpha_0 + \frac{\lambda}{b}\right). \end{aligned}$$

Example: $\alpha = 30° \lambda/b = 0.2$

$$\Rightarrow \Delta\theta = 44.4° - 17.6° \approx 26.8,$$

For $\alpha_0 = 0_0$ is $\Delta\theta_0 = 25.6°$.

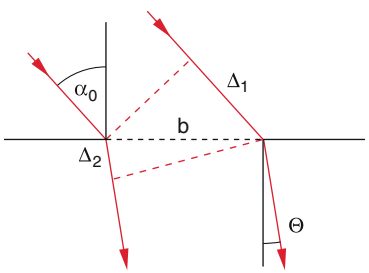


Fig. A.44 Illustration to solution 10.5

10.6 (a) From the grating equation

$$d \cdot (\sin\alpha + \sin\beta) = m \cdot \lambda$$

we get for $m = 1$ and $\alpha = 30°$.
The diffraction angle is on the opposite side of the incidence angle α. The angle between diffracted beam and incident beam is $\Delta\varphi = \alpha - \beta = 31.5°$.
For $m = 2$ is

$$\sin\beta^{(2)} = 2\frac{\lambda}{b} - \sin\alpha = 0.96 - 0.5 = 0.46$$

The second order diffraction appears at the angle $\beta_2 = 14.35°$.

(b) the blaze angle is

$$\theta = \frac{\alpha+\beta}{2} = \frac{30 - 1.3}{2} = 14.35°.$$

(c) The angular difference $\Delta\beta$ between the first and the second diffraction order can be obtained from

$$\sin\beta_1 - \sin\beta_2 = \frac{\lambda_1 - \lambda_2}{d} = \frac{-10^{-9}\,\text{m}}{-10^{-6}\,\text{m}}$$

(d) The lateral distance between the centers of the two slit images $b(\lambda_1)$ and $b(\lambda_2)$ is

$$\Delta b = f \cdot \Delta\beta = 1\,\text{mm}.$$

for a grating with area 10 × 10 cm is the diffraction limited basis width

$$\begin{aligned}\Delta b &= 2 \cdot \frac{\lambda}{d} \cdot f \\ &= 2 \cdot \frac{4.8 \times 10^{-7}\,\text{m}}{10^{-2}\,\text{m}} \cdot 1\,\text{m} = 9.6 \times 10^{-5}\,\text{m} \\ &\approx 0.1\,\text{mm}.\end{aligned}$$

The width of the entrance slit should be no larger than 0.9 mm.

10.7 According to (10.9) the phase difference between the partial waves reflected at the interfaces air-oil and oil-water including the phase jump is

$$\Delta\varphi = \frac{2\pi}{\lambda_0}\Delta s - \pi.$$

For constructive interference the phase difference has to be $\Delta\varphi = 2m \cdot \pi$. The path difference is then

$$\Rightarrow \Delta s = \frac{2m+1}{2}\lambda_0.$$

Since $\Delta s = 2d \cdot \sqrt{n^2 - \sin^2\alpha}$ we get for $\lambda = 500\,\text{nm}$ (green)

$$d = \frac{\Delta s}{\sqrt{n^2 - \sin^2\alpha}} = \frac{(m+1/2)\lambda_0}{\sqrt{n^2 - \sin^2\alpha}}.$$

For $m = 0$, i.e. $\alpha = 45°$ is the numerical value

$$\begin{aligned}d &= \frac{2.5 \times 10^{-7}\,\text{m}}{\sqrt{1.6^2 - 0.5}} = 1.74 \times 10^{-7}\,\text{m} \\ &= 0.174\,\mu\text{m}.\end{aligned}$$

10.8 The distance between the tilted planes with the wedge angle ε is (Figs. A.44 and A.45)

$$d(x) = x \cdot \tan\varepsilon.$$

For a sufficiently small angle ε we can neglect the small inclination angle of the light reflected at the lower surface. The thickness of the glass plates is assumed to be large compared to that of the air wedge and also larger than the coherence length of the incident light. This implies that the glass plates do not act as interferometers and do not cause additional interferences.
Constructive interference between the light reflected at the upper and the lower interface of the air wedge appears w if the phase difference is

$$\Delta\varphi = \frac{2\pi}{\lambda_0}\Delta s - \pi = 2m \cdot \pi$$

(Note the phase jump for the reflection by the lower edge of the wedge).
With $\Delta s = 2d(x) = 2x \cdot \tan\varepsilon$ we get

$$2x \cdot \tan\varepsilon = \left(m + \frac{1}{2}\right)\lambda.$$

With the distance Δx between the interference stripes we get with $2\Delta x \tan\varepsilon = \lambda$ for $m = 1$.

$$\begin{aligned}&\Rightarrow \tan\varepsilon = \frac{\lambda}{2\Delta x} = \frac{5.89 \times 10^{-7}}{2 \cdot \frac{1}{12} \times 10^{-2}} = 3.5 \times 10^{-4} \\ &\Rightarrow \varepsilon = 0.02°\end{aligned}$$

10.9 With the amplitude A_0 of the light transmitted by the smaller slit with width b the transmitted intensity is

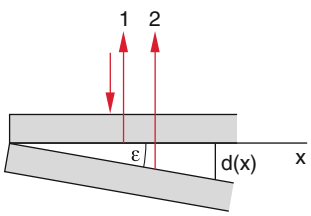


Fig. A.45 Illustration to solution 10.8

$$I_0 = c \cdot \varepsilon_0 A_0^2$$

The intensity transmitted by the larger slit with the twofold width 2b is $2I_0$, the amplitude is therefore $\sqrt{2}A_0$. The total amplitude at the point P (Fig. A.46) is therefore

$$I = c \cdot \varepsilon_0 \cdot \left| A_0 + \sqrt{2}A_0 \cdot \mathrm{e}^{\mathrm{i}\Delta\varphi} \right|^2,$$

where

$$\Delta\varphi = \frac{2\pi}{\lambda} \cdot \Delta s = \frac{2\pi}{\lambda} d \cdot \sin\theta$$

is the phase difference between the two partial waves and d is the distance between the two slits. This gives the total intensity

$$\begin{aligned} I &= I_0 \cdot \left[\left(1 + \sqrt{2}\,\mathrm{e}^{\mathrm{i}\,\Delta\varphi}\right)\left(1 + \sqrt{2}\mathrm{e}^{-\mathrm{i}\,\Delta\varphi}\right) \right] \\ &= I_0 \cdot \left(3 + 2 \cdot \sqrt{2}\,\cos\,\Delta\varphi\right) \\ &\Rightarrow I_{\max} = 5.83\, I_0 \\ I_{\min} &= 0.172\, I_0. \end{aligned}$$

10.10 The first zero of the function $\sin^2 x/x^2$ appears at $x = \pi$, the second at $x = 2\pi$.
The first maximum is found by setting the first derivative equal to zero.

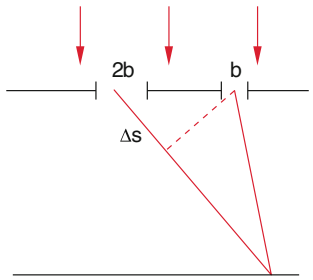


Fig. A.46 Illustration to solution 10.9

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}x}\left(\frac{\sin^2 x}{x^2}\right) = 2\left(\frac{x\,\cos x}{x^2} - \frac{\sin x}{x^2}\right) \\ &\Rightarrow x \cdot \cos x = \sin x \Rightarrow x = \tan x \\ &\Rightarrow x = 4.4934 = 1.43\pi. \end{aligned}$$

The relative deviation from the value 1.5π between the two maxima is therefore

$$\Delta = \frac{1.5 - 1.43}{1.5} = 4.67\%.$$

10.11 The angular width between the two basis points $\pm\theta_1$ of the central diffraction maximum is, according to (10.46), with $\sin\theta_1 = \pm 1.2 \cdot \lambda/D$ and $\theta_1 << 1$

$$\Delta\theta = 2.4 \cdot \lambda/D.$$

(a) The mean distance to the moon is $r = 3.8 \times 10^8$ m. The diameter of the central diffraction maximum on the moon is then

$$\begin{aligned} d = r \cdot \Delta\theta &= 3.8 \times 10^8 \cdot 2.4 \cdot \frac{6 \times 10^{-7}}{1}\ \mathrm{m} \\ &= 5.47 \times 10^2\ \mathrm{m}. \end{aligned}$$

(b) The retro-reflector with area A on the moon receives the fraction

$$\varepsilon_1 = \frac{A}{\pi(d/2)^2} = \frac{0.25}{\pi \cdot 2.7^2} \times 10^{-4} \approx 10^{-6}$$

of the emitted intensity. The radiation reflected by the reflector has the diffraction angle

$$\Delta\theta_2 = \frac{\lambda}{0.5\ \mathrm{m}} = 1.2 \times 10^{-6}.$$

This reflected light covers on the earth surface an area

$$A_2 = (r \cdot 1.2 \times 10^{-6})^2 = (3.8 \cdot 1.2 \times 10^2)^2\ \mathrm{m}^2.$$

The telescope receives the fraction

$$\varepsilon_2 = \frac{\pi(D/2)^2}{A^2} = 3.8 \times 10^{-6}.$$

of this reflected light intensity.

(c) Without retroreflector about 30% of the total intensity received by the moon, would be backscattered into the solid angle $\Omega = 2\pi$. The telescope would receive the fraction

$$\begin{aligned} \varepsilon_3 &= \frac{\pi(D/2)^2}{2\pi \cdot r} = \frac{D^2}{8r^2} = \frac{1}{8 \cdot 3.8^2 \times 10^{16}} \\ &= 8.6 \times 10^{-19} \end{aligned}$$

of this back-scattered light. The retroreflector therefore increases the intensity received by the telescope by the factor

$$\frac{\varepsilon_1 \cdot \varepsilon_2}{0.3 \cdot \varepsilon_3} = \frac{3.8 \times 10^{-12}}{0.3 \cdot 8.6 \times 10^{-19}} = 1.5 \times 10^7!$$

10.12 (a) The refractive index n_1 of the antireflection coating may be larger or smaller than the index n_2 of the substrate. If it is larger, the incident wave suffers a phase jump at the first interface air-coating, but not at the second one coating-substrate. For destructive interference of

the waves reflected by the two interfaces the layer thickness must be

$$d = m \cdot \lambda_0/(2n_1)$$

If n_1 is smaller than n_2 the wave reflected at the interface coating-substrate suffers a phase jump of π. The anti-refection coating must now have a thickness of

$$d = \frac{2m+1}{4}\frac{\lambda_0}{n_1}$$

The total amplitude of the reflected wave is

$$\begin{aligned} A &= \sqrt{R_1}A_0 - (1-R_1)\sqrt{R_2}A_0 + (1-R_1) \\ &\quad \times R_2\sqrt{R_1}A_0 - (1-R_1)R_2^{3/2}R_1A_0 - \cdots \\ &= A_0\sqrt{R_1} - (1-R_1)\sqrt{R_2}(1+\sqrt{R_1R_2}) \\ &\quad + R_1R_2 + (R_1R_2)^{3/2} + \cdots \\ &= A_0\left[\sqrt{R_1} - (1-R_1)\sqrt{R_2}\cdot\frac{1}{1-\sqrt{R_1R_2}}\right] \\ &= A_0\left(\frac{\sqrt{R_1}-\sqrt{R_2}}{1-\sqrt{R_1R_2}}\right). \end{aligned}$$

The amplitude becomes minimum for $\sqrt{R_1} = \sqrt{R_2}$. This implies

$$\begin{aligned} &\frac{n_1 - n_{\text{Luft}}}{n_1 + n_{\text{Luft}}} = \frac{n_2 - n_1}{n_2 + n_1} \\ &\Rightarrow n_1^2 = n_{\text{Luft}}n_2. \end{aligned}$$

The above considerations show that

$$d = \lambda/4 + m \cdot \lambda/2 (m = 0, 1, 2, \ldots)$$

(b) (Fig. A.47) One has the solve the equation

$$\sqrt{R_1}A_0 - (1-R_1)\sqrt{R_2}A_0 - (1-R_1)\sqrt{R_1}A_0 = 0$$

for n_1. This represents an equation of third order, which can be best solved with a computer program. The result is for $n_{\text{Luft}} = 1$ and $n_2 = 1.5$

$$n_1 = 1.22473198\ldots,$$

which deviates from the experimental value only by 0.001%.

10.13 If the hexagon is supplemented by an equilateral triangle (Fig. A.48) with the apex angle $\gamma = 60°$, the situation in Sect. 9.4 is copied. There it was shown that the total deflection of the incident beam is

$$\delta = (\alpha_1 - \beta_1) + (\alpha_2 - \beta_2)$$

The minimum deflection $\delta_{\min}$ is obtained for

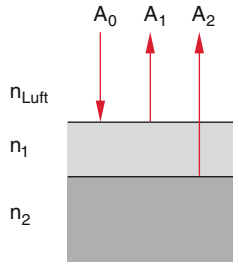


Fig. A.47 Illustration to solution 10.12b

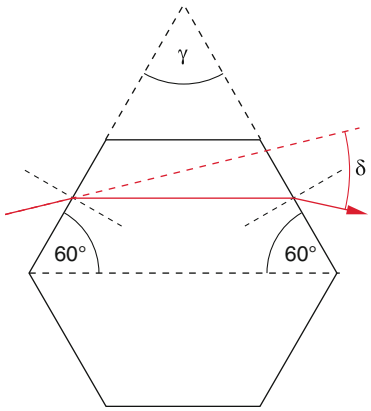


Fig. A.48 Illustration to solution 10.13

$$\begin{aligned} &\alpha_1 = \alpha_2 = \alpha \quad \text{and} \quad \beta_1 = \beta_2 = \beta \\ &\gamma = 2\beta \rightarrow \delta_{\min} = 2\alpha - \gamma \end{aligned}$$

From Snellius rule $\sin\alpha/\sin\beta = n \Rightarrow$

$$\begin{aligned} &\sin\left(\frac{\delta_{\min}+\gamma}{2}\right) = n \cdot \sin\beta = n \cdot \sin(\gamma - 2) \\ &\Rightarrow \sqrt{3} \cdot \sin\left(\frac{\delta_{\min}}{2}\right) + \cos\left(\frac{\delta_{\min}}{2}\right) = n = 1.31 \\ &\Rightarrow \delta_{\min} = 22°. \end{aligned}$$

10.14 The scattering cross section is

$$\begin{aligned} &\sigma_s = a\frac{\omega^4}{(\omega_0^2-\omega^2)^2 + \omega^2\gamma^2} \\ &\frac{\mathrm{d}\sigma}{\mathrm{d}\omega} = 0 = a \cdot \left\{\frac{4\omega^2\left[(\omega_0^2-\omega^2)^2 + \omega^2\gamma^2\right]}{N^2}\right. \\ &\qquad \left. - \frac{\omega^4\left[-4\omega(\omega_0^2-\omega^2) + 2\gamma^2\omega\right]}{N^2}\right\} \\ &\Rightarrow (\omega_0^2-\omega^2)^2 + \omega^2\gamma^2 + \omega^2(\omega_0^2-\omega^2) \\ &\qquad - \frac{1}{2}\gamma^2\omega^2 = 0 \\ &\Rightarrow \omega_{\text{m}} = \frac{\omega_0^2}{\sqrt{\omega_0^2 - \gamma^2/2}} = \frac{\omega_0}{\sqrt{1-\gamma^2/2\omega_0^2}}. \end{aligned}$$

Chapter 11

11.1 The imaging equation is

$$\frac{1}{a}+\frac{1}{b}=\frac{1}{f}.$$

Since here is $a \gg b$ we can approximate $f \approx b = 2\,\text{m}$. The diameter of the sun image is

$$d=\frac{b}{a}\cdot D=\frac{2}{1.5\times 10^{11}}\cdot 1.5\times 10^{9}\,\text{m} = 2.10^{-2}\,\text{m}=2\,\text{cm}.$$

With the naked eye the sun appears under the angle

$$\varepsilon_0\frac{D}{r}=\frac{1.5\times 10^{9}}{1.5\times 10^{11}}=10^{-2}\,\text{rad}\approx 0.5^\circ$$

if the sun image generated by the lens is observed at the distinct visual range $s_0 = 25\,\text{cm}$ the visual angle is

$$\varepsilon=\frac{2}{25}=8\times 10^{-2}\,\text{rad}.$$

The angular magnification is therefore V = 8. The lateral reduction is

$$V=\frac{b}{a}=\frac{2}{1.5\times 10^{11}}=1.3\times 10^{-11}.$$

11.2 The magnification of the visual angle is according to (11.45)

$$V_\text{L}=\frac{s_0}{f}\left(1+\frac{f-g}{g}\right)=\frac{25}{2}\left(1+\frac{0.5}{1.5}\right) = 16.7.$$

From

$$\frac{1}{f}=\frac{1}{g}+\frac{1}{b}$$

follows

$$b=\frac{g\cdot f}{g-f}=\frac{3}{0.5}\text{cm}=-6\,\text{cm}.$$

From Fig. 11.8 one can see, that the virtual image of the letter G is with $B/G = -b/g$

$$B=-G\cdot\frac{b}{g}=0.5\,\frac{6}{1.5}\,\text{mm}=2\,\text{mm}$$

The lateral magnification is therefore fourfold.

11.3 Different from the derivation of Eq. (9.26) the derivation of (11.2) demands the consideration of the three different refractive indices n_1, n_2 and n_3. Equation (9.23a) then becomes

$$\frac{n_1}{g_1}+\frac{n_\text{L}}{b_1}=\frac{n_\text{L}-n_1}{R_1},$$

and (9.23b) changes to

$$-\frac{n_\text{L}}{b_1-d}+\frac{n_2}{b_2}=\frac{n_2-n_\text{L}}{R_2}.$$

After addition according to (9.24a) and the approximation (9.24b) for thin lenses one obtains

$$\frac{n_1}{g}+\frac{n_2}{b}=\frac{n_\text{L}-n_1}{R_1}-\frac{n_\text{L}-n_2}{R_2}\overset{\text{def}}{=}X. \qquad (*)$$

For $g=\infty$ is $b = f_2$ and we get

$$\frac{n_1}{f_2}=\frac{n_\text{L}-n_1}{R_1}-\frac{n_\text{L}-n_2}{R_2}=X.$$

In a similar way one finds

$$\frac{n_1}{f_1}=X.$$

These two equations can be rearranged into

$$n_1=f_1\cdot X \text{ and } n_2=f_2\cdot X.$$

If this is inserted into Eq. (*) one obtains after dividing by X

$$\frac{f_1}{g}+\frac{f_2}{b}=1.$$

11.4 According to (11.8b) is

$$\delta_\text{min}=1.22\frac{\lambda}{D}<\varepsilon=1.5''=7.2\times 10^{-6}\text{rad}$$
$$\Rightarrow D>\frac{1.22\cdot\lambda}{\varepsilon}=0.084\,\text{m}=8.4\,\text{cm}.$$

The diameter of the pupil is at night about 5 mm. The eye has its maximum sensitivity at $\lambda = 500\,\text{nm}$

$$\Rightarrow \varepsilon_\text{min}=\frac{1.22\cdot\lambda}{D}=1.22\times 10^{-4}\text{rad}=25''.$$

11.5 The diameter of Jupiter is 71,398 km. The radius of its orbit around the sun is r = 5.2 AU (astronomical units). At its closest approach to earth its distance to earth is

$$\Delta r=(5.2-1)\,\text{AE}=4.2\text{AE}=6.3\times 10^{11}\,\text{m}.$$

For the naked eye its diameter appears under the angle

$$\varepsilon_0=\frac{7.14\times 10^{7}}{6.3\times 10^{11}}=1.13\times 10^{-4}\text{rad}=23''$$

This angle is large compared to the random variations $\Delta\varepsilon \approx 1''$ caused by air turbulences. The image of Jupiter therefore does not fluctuate like the twinkling stars. The same is true for Venus and Mars.

11.6 The angle under which the diameter of a tennis ball is observed by the satellite, is

$$\varepsilon = \frac{d}{r} \approx \frac{10^{-1}}{4.5^5}\text{rad} = 2.5 \times 10^{-7}\text{rad} = 0.05''$$

The mirror of the telescope should have a diameter

$$D = \frac{1.22 \cdot \lambda}{\varepsilon} = \frac{1.22 \cdot 4 \times 10^{-7}}{2.5 \times 10^{-7}} \approx 2\,\text{m}$$

Because of the air turbulence which limits the angular resolution to about 1" the smallest dimension on earth which can be resolved by the satellite is about 2 m. With special techniques of image processing this limit can be improved by about a factor of 4. This allows the resolution of 50 cm with a telescope with 1 m diameter on board of the satellite.

$$\begin{aligned}\delta_{\min} = \frac{\Delta x}{r} &= \frac{1}{10^4} = 10^{-4}\text{rad} \\ \Rightarrow D &= \frac{1.22\lambda}{\delta_{\min}} = \frac{1.22 \cdot 0.01}{10^{-4}}\,\text{m} \\ &= 1.22 \times 10^2\,\text{m} = 122\,\text{m}!\end{aligned}$$

This cannot be reached with a single antenna but is realized with an array of antennas, which have distances of about 100 m.

11.8 The magnification of the microscope must be 50 times.

$$\varepsilon_0 = \frac{D_0}{s_0} = \frac{2 \times 10^{-5}}{0.25} = 8 \times 10^{-5}.$$

The objective lens brings the angular magnification

$$\frac{\varepsilon_1}{\varepsilon_0} = V_1 = 10 \Rightarrow \varepsilon_1 = 8 \times 10^{-4}.$$

With $\varepsilon_1 = D_0/g$ is follows

$$g = \frac{D_0}{\varepsilon_0} = \frac{2.10^{-5}}{8.10^{-4}}\,\text{m} = 2.5 \times 10^{-2}\,\text{m} = 2.5\,\text{cm}.$$

We choose the focal length $f_1 = 2\,\text{cm}$

$$\Rightarrow b = \frac{gf_1}{g - f_1} = \frac{2.5 \cdot 2}{0.5}\text{cm} = 10\,\text{cm}.$$

The total magnification of the microscope is

$$\begin{aligned}V_{\text{M}} &= \frac{b \cdot s_0}{gf_2} \\ \Rightarrow f_2 &= \frac{b \cdot s_0}{g \cdot V_{\text{M}}} = \frac{10.25}{2.5 \cdot 50}\,\text{cm} = 2\,\text{cm}.\end{aligned}$$

11.9 We start with the grating equation

$$d \cdot (\sin\alpha + sin\beta) = m \cdot \lambda.$$

For $m = 1$ one gets for the difference of the diffraction angles β_1 and β_2 the condition

$$\sin\beta_1 - \sin\beta_2 = \frac{\lambda_1}{d} - \frac{\lambda_2}{d}.$$

The distance between the two slit images is

$$\begin{aligned}\Delta x_{\text{B}} &= f_2 \cdot \left(\frac{\lambda_1}{d} - \frac{\lambda_2}{d}\right) \\ &= \frac{3}{10^{-6}}(501 - 500) \times 10^{-9}\,\text{m} \\ &= 3 \times 10^{-3}\,\text{m} \;\; = 1\,\text{mm}.\end{aligned}$$

The angular width between the two zeros on both sides of the central diffraction maximum is

$$\Delta\alpha = \frac{2\lambda}{D} \Rightarrow \Delta x = f_2 \cdot \frac{2\lambda}{D}.$$

With $D = 10$ cm we get

$$\Delta x = \frac{2.5 \times 10^{-7} \cdot 3}{0.1} = 3 \times 10^{-5}\,\text{m} = 30\,\mu\text{m}.$$

For the slit width b the total with of the slit image becomes

$$\Delta x_{tot} = b + 30\,\mu\text{m} \leq 1\,\text{mm}.$$

Therefore b has to be smaller than 0.97 mm in order to achieve the complete resolution of the two spectral lines.

11.10 The free spectral range of the Fabry-Perot Interferometer (FPI) is according to (10.28)
With n = 1 (air spaced FPI) and d = 1 cm we get

$$\delta\nu = 1.5 \times 10^{10}\text{s}^{-1} = 15\,\text{GHz}.$$

Expressed in wavelengths this becomes

$$\delta\lambda = -\frac{\lambda^2}{c}\delta\nu = -\frac{\lambda^2}{2nd}.$$

For $\lambda = 500$ nm is

$$\delta\lambda = -\frac{25 \times 10^{-14}}{2 \times 10^{-2}} = 12.5 \times 10^{-2}\,\text{m} = 12.5\,\text{pm}.$$

The finesse is

$$F^* = \frac{\pi \cdot \sqrt{R}}{1-R} = \frac{\pi \cdot \sqrt{0.98}}{0.02} = 155.$$

If the FPI plates are ideals planes and the alignment is perfect the spectral resolving power is

$$\left|\frac{\lambda}{\Delta\lambda}\right| = \left|\frac{\nu}{\Delta\nu}\right| = \frac{\Delta s_m}{\lambda} = F^* \cdot \frac{2d}{\lambda}$$
$$= 155 \cdot \frac{2 \times 10^{-2}}{5 \times 10^{-7}} = 6.2 \times 10^6$$

This means that two wavelengths with a separation

$$\Delta\lambda = \frac{\lambda}{6.2 \times 10^6} = \frac{5 \times 10^{-7}}{6.2 \times 10^6}\,\text{m}$$
$$= 8 \times 10^{-14}\,\text{m} = 0.08\,\text{pm}$$

can be still resolved.
Measured in frequencies ν we get

$$\Delta\nu = \delta\nu/F^* = \frac{15\,\text{GHz}}{155}$$

Note: Here we have used a finesse F^* = 155 and the relation
$\Delta\nu = \delta\nu/F^*$ between spectral width $\Delta\nu$ of a spectral line and the free spectral range $\delta\nu$ of the FPI.

(b) The free spectral range of the FPI in units of the wavelength is

$$\delta\lambda = 12.5\,\text{pm} = 0.0125\,\text{nm}.$$

The lateral distance of the slit images of two spectral lines in a spectrograph with the focal length f_2 with this wavelength separation $\delta\lambda$ is

$$\Delta x = f_2 \cdot \frac{dn}{d\lambda}\delta\lambda > b$$

$\Longrightarrow$ The focal length has to be at least

$$\Rightarrow f > \frac{b}{dn/d\lambda \cdot \delta\lambda} = \frac{10^{-5}}{5 \times 10^5 \cdot 12.5 \times 10^{-12}}\,\text{m}$$
$$= 1.6\,\text{m}$$

Chapter 12

12.1 (a) The distance between lens and focal plane can be obtained from the lens equation

$$\frac{1}{a} + \frac{1}{b} = \frac{1}{f} \Rightarrow b = \frac{a \cdot f}{a - f} = \frac{100 \times 10}{90}\,\text{mm} = 11.1\,\text{mm}.$$

The focus diameter d_2 is according to geometrical optics given by

$$d_2 = \frac{b}{a} \cdot d_1 = \frac{11.1}{100} \cdot 0.01\,\text{mm} = 1.1\,\mu\text{m}.$$

Taking into account diffraction the distance between the two zeros on both sides of the central diffraction maximum is for a lens diameter D and parallel incident light

$$d_2 = 2.4(\lambda/D) \cdot f = 2.4 \cdot (600 \times 10^{-6}/5) \times 10\,\text{mm}$$
$$= 2.88 \times 10^{-3}\,\text{mm} = 2.88\,\mu\text{m}.$$

For a point like source the diffraction limited image is already larger than the geometrical image for an aperture diameter of 10 µm.
The actual image in the focal plane is the convolution of the geometrical image and the diffraction intensity distribution.

(b) According to geometrical optics the Rayleigh length Δz is the distance between the focal plane at z = 0 and the plane z = Δz where the diameter of the divergent beam has enlarged by $\sqrt{2}$.

Using the intercept theorem is (Fig. A.49)

$$\frac{D/2 - d_2}{b} = \frac{(\sqrt{2}-1)d_2}{\Delta z}$$
$$\Rightarrow \Delta z = \frac{b \cdot d_2(\sqrt{2}-1)}{D/2 - d_2} \approx \frac{0.414b \cdot d_2}{D/2}$$
$$\approx \frac{0.414.11, 1.1, 1.10^{-3}}{2.5}\,\text{mm} = 2\,\mu\text{m}.$$

The exact calculation using Gauss-profiles for the light beam gives

$$\Delta z = \pi \cdot d_2^2/\lambda = \pi \cdot 1.1^2 \times 10^{-6}/6 \times 10^{-4}\,\text{mm} = 2\,\mu\text{m},$$

which is identical to the result of geometrical optics.
The numerical values illustrate that the focus depth of confocal microscopy is about 1 µm.

12.2 (a) The angular range $2\Delta\theta$ between the two zeros at both sides of the central maximum is for a mirror diameter D

$$2\Delta\vartheta = 2 \cdot 1.2\lambda/D = 2.4 \cdot 5 \times 10^{-7}/5$$
$$= 2.4 \times 10^{-7}\,\text{rad} \triangleq 0.051''.$$

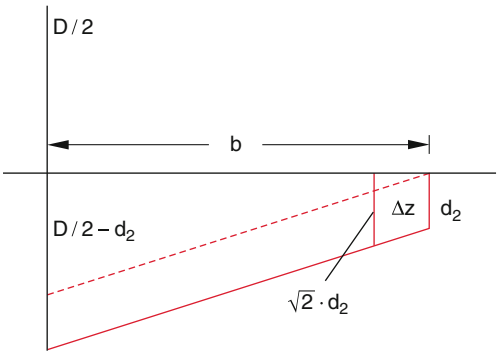


Fig. A.49 Illustration to solution 12.1

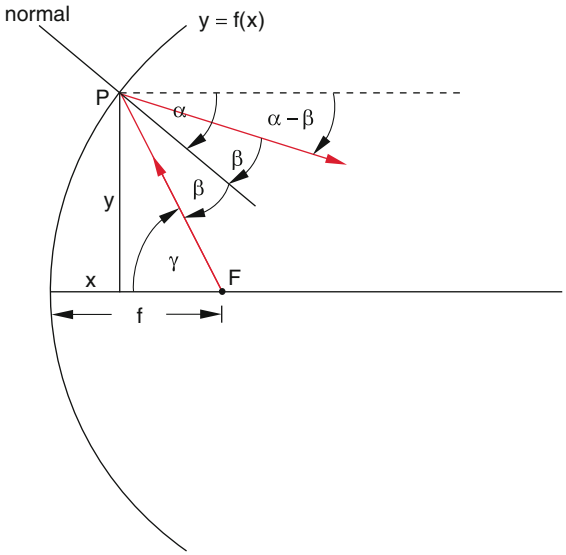


Fig. A.50 Illustration to solution 12.2

The diameter d for a focal length f is then

$$d = f \cdot 2\Delta\vartheta = 2.4 \times 10^{-7} \times 10\,\text{m} \hat{=} 2.4 \times 10^{-6}\,\text{m} = 2.4\,\mu\text{m}.$$

(b) We regard a ray in Fig. A.50 which starts from the focus F and hist the mirror at $P(x, y)$ where it is reflected. For the ideal parabolic mirror all reflected rays should be parallel. For the deformed mirror the direction $(\alpha - \beta)$ of the reflected ray depends on the point $P(x, y)$. It is

$$y^2 = 4f\,\varepsilon x \Rightarrow \frac{dy}{dx} = \frac{f \cdot \varepsilon}{\sqrt{f\varepsilon x}} \Rightarrow \frac{dy}{dx} = \sqrt{x/f\varepsilon}.$$

The slope of the normal to the mirror surface in the point $P(x, y)$ is obtained from

$$-\tan\alpha = \frac{dx}{dy} = \sqrt{x/f\varepsilon} \Rightarrow \tan\alpha = -\frac{y}{2f\varepsilon}.$$

From Fig. A.50 we see that

$$\tan\gamma = \frac{y}{f - x}.$$

The slope of the reflected beam is $-\tan(\alpha - \beta)$. With the relation

$$\tan(\alpha - \beta) = \frac{\tan\alpha - \tan\beta}{1 + \tan\alpha \cdot \tan\beta},$$

$$\gamma = -(\alpha + \beta) \Rightarrow \beta = -(\alpha + \gamma),$$

$$\tan\beta = -\frac{\tan\alpha + \tan\gamma}{1 - \tan\alpha \cdot \tan\gamma}$$

$$\Rightarrow \tan(\alpha - \beta) = \frac{2\tan\alpha + \tan\gamma(1 - \tan^2\alpha)}{1 - 2\tan\alpha \cdot \tan\gamma - \tan^2\alpha}.$$

we obtain

$$\tan\beta = -\frac{\tan\alpha + \tan\gamma}{1 - \tan\alpha \cdot \tan\gamma}$$

$$\Rightarrow \tan(\alpha - \beta) = \frac{2\tan\alpha + \tan\gamma(1 - \tan^2\alpha)}{1 - 2\tan\alpha \cdot \tan\gamma - \tan^2\gamma}$$

Inserting $\tan\alpha$ and $\tan\gamma$ yields

$$\tan(\alpha - \beta) = \frac{f \cdot y(\varepsilon - 1)}{f^2\varepsilon + 3xf\varepsilon - xf + x^2}.$$

For $\varepsilon = 1$ is $(\alpha - \beta) = 0$, i.e. the reflected beam is always parallel to the symmetry axis. For $\varepsilon > 1$ the maximum deviation from the parallel direction $(\alpha - \beta) = 0$ is reached for $x = D/2$, i.e. at the edge of the mirror, where

$$x = \frac{y^2}{4f\,\varepsilon} = \frac{D^2}{16f \cdot \varepsilon}.$$

Inserting into the relation for $\tan(\alpha - \beta)$ gives

$$\tan(\alpha - \beta)_{\max} = \frac{(f/D)(\varepsilon - 1)}{(f/D)^2\varepsilon + \dfrac{3\varepsilon - 1}{16\varepsilon} + \dfrac{D^2}{16^2 f^2 \varepsilon^2}}.$$

For $\varepsilon = 1.01$ and $f = 4D \Rightarrow$ we get

$$\tan(\alpha - \beta)_{\max} = \frac{0.04}{16.16 + \dfrac{2.03}{16.16} + \dfrac{1}{16^3}} \approx 0.0025$$

The maximum inclination angle of the reflected ray is

$$\Delta(\alpha - \beta) = 0.0025\,\text{rad} = 0.15^0.$$

For $\varepsilon = 1.001$ (0.1% deviation from the ideal parabolic mirror)

$$\Rightarrow \tan(\alpha - \beta)_{\max} \approx \frac{0.004}{16.02} = 2.5 \times 10^{-4}$$

$$\Rightarrow (\alpha - \beta)_{\max} = 53''!$$

Even such a small deformation of the parabolic mirror would deteriorate the angular resolution of the mirror about 50 times more than the air turbulence (seeing).

12.3 We have discussed in Sect. 9.7 that the difference

$$\varrho = \zeta_w - \zeta_s \approx a(n_0 - 1) \cdot \tan\zeta$$

between real and apparent zenith distance ζ is determined experimentally as

$$\varrho \approx 58.2'' \cdot \tan\zeta$$

The relative density fluctuation $\delta n/n$ of the atmosphere causes a smear

$$\delta\varrho = \frac{\delta n}{n} \cdot a \cdot (n_0 - 1)\tan\zeta = \frac{\delta n}{n} \cdot 58.2'' \tan\zeta$$
$$= 3 \times 10^{-2} \cdot 58.2'' \tan\zeta = 3.0''.$$

12.4 The central diffraction maximum at a groove width of $b = 2\,\mu\text{m}$ lies in the angular range

$$\frac{-\lambda}{b} \leq \sin\alpha \leq +\frac{\lambda}{b} \Rightarrow |\sin\alpha| \leq \frac{\lambda}{b} = 0.25$$
$$\Rightarrow |\alpha| \leq 14.48°.$$

The interference maxima appear according to (12.33) at the angles α determined by

$$\sin\alpha = \frac{2b}{d}(n-1) - (m_2 - m_1) \cdot \frac{2\lambda}{d}.$$

Possible angles α_i with $|\alpha_i| \leq 14.48°$ are

(1) $m_2 - m_1 = 0$: $\Rightarrow \sin\alpha_0 = 0.2 \Rightarrow \alpha_0 = 11.5°$
(2) $m_2 - m_1 = 1$: $\Rightarrow \sin\alpha_1 = -0.05 \Rightarrow \alpha_1 = -2.87°$.

For $m_2 - m_1 = -2$: $\Rightarrow \sin\alpha_2 = -0.3 \Rightarrow \alpha_2 = -17.4°$. This is already outside the central diffraction maximum and therefore only a very small intensity is found for this interference maximum. The same is true for all other interference orders.

(b) For an incidence angle $\alpha_e \neq 0$ the condition for the zero at both sides of the central diffraction maximum is with $\alpha_0 = \alpha_e$

$$b(\sin\alpha_e - \sin\alpha) = \pm\lambda.$$

this gives with $\sin\alpha_e = 0.5$:

$$\sin\alpha = 0.25 \text{ or} 0.75$$
$$\Rightarrow 14.48° \leq \alpha \leq 48.6°.$$

The condition (12.33) is now

$$(n-1)b - \frac{d}{2}(\sin\alpha - \sin\alpha_e) = (m_2 - m_1)\lambda$$
$$\Rightarrow \sin\alpha = \frac{2b}{d}(n-1) - \frac{2\lambda}{d}(m_2 - m_1) + \sin\alpha_e.$$

This gives for $m_2 - m_1 = 0$:

$$\sin\alpha_0 = \frac{2}{4} \cdot 0.4 + 0.5 = 0.7 \Rightarrow \alpha_0 = 44.4°$$

and for $m_2 - m_1 = +1$

$$\sin\alpha_1 = 0.45 \Rightarrow \alpha_1 = 26.7°.$$

All other interference orders lie outside the central diffraction maximum

12.5 (a) The depth of the grooves for $\lambda = 600\,\text{nm}$ can be obtained from the condition

$$(n-1) \cdot h = \lambda/2 \Rightarrow h = (0.3/0.5)\,\mu\text{m} = 0.6\,\mu\text{m}.$$

The radii of the rings are

$$r_m = \sqrt{m \cdot s_0 \cdot \lambda}.$$

The focal length is $f = s_0 = r_1^2/\lambda = 10\,\text{mn}$

$$\Rightarrow r_1 = \sqrt{f \cdot \lambda} = 7.7 \times 10^{-5}\,\text{m} = 77\,\mu\text{m}.$$

The maximum radius of the outermost ring is $r_m = d/2 = 10^{-2}\,\text{m} = \sqrt{m \cdot f \cdot \lambda} = 77\sqrt{m}\,\mu\text{m}$

$$\Rightarrow m = r_m^2/(f \cdot \lambda) = \frac{10^{-4}}{10^{-2} \cdot 6 \times 10^{-7}} = 1667.$$

(b) The focal length of a refractive biconvex lens with radii of curvature $R_1 = R_2 = R$ is

$$f = \frac{1}{n-1}\frac{R}{2}.$$

The diameter D is for a given radius of curvature R maximum for a sphere, where $D_{\max} = 2R$.

$$\Rightarrow f = \frac{1}{n-1} \cdot \frac{D_{\max}}{4} \Rightarrow D_{\max} = 4(n-1) \cdot f.$$

For $n = 1.5 \Rightarrow D_{\max} = 2f$.

The aberrations of such a lens are very large. For a plane-convex half-sphere is $D_{\max} = f$.

(c) A Fresnel zone plate can be produced in the following way (Fig. A.51): A plane wave is superimposed with a spherical wave. At the distance s_0 from the center of the spherical wave a photo-plate is positioned. The interference pattern in the plane of the photo-plate corresponds to the Fresnel-zone arrangement. The photo-plate is developed and then

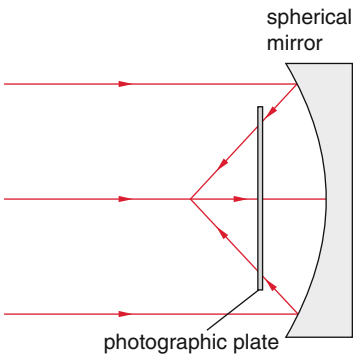


Fig. A.51 Illustration to solution 12.5

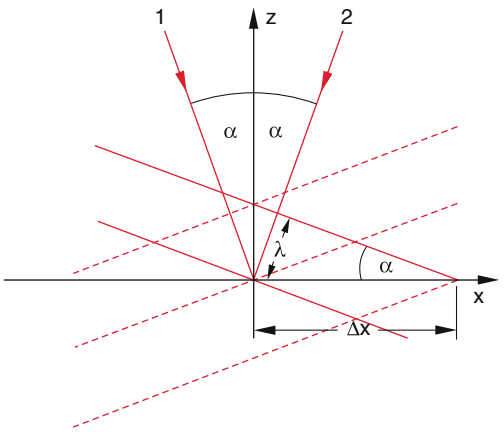


Fig. A.52 Illustration to solution 12.6

circular ring grooves are produced by chemical etching techniques.

12.6 If two plane waves propagating in the x-z-plane impinge onto the x-y-plane under the angles $\pm\alpha$ against the normal to the plane (Fig. A.52) the path difference $\Delta s = s_1 - s_2$ between the two waves on the line $x = 0$ (i.e. in the y-direction) is zero.
Along the x-direction the path length s_1 changes by $\Delta s_1 = +\Delta x$, for s_2 by $\Delta s_2 = -\Delta x \cdot \sin\alpha \Rightarrow \Delta s = 2\Delta x \cdot \sin\alpha \Rightarrow$. The path difference between the two waves is then.
Interference maxima appear in the x-direction for $\Delta s = m \cdot \lambda$. The distance between the lines x = const of maximum intensity is therefore

$$\Delta x = \frac{\lambda}{2 \cdot \sin\alpha}.$$

Example: In order to achieve a distance $\Delta x = 1\,\mu\text{m}$ under illumination with $\lambda = 500\,\text{nm}$, we get $\sin\alpha = 0.5/2 = 0.25\,\text{sein}, \Rightarrow \alpha = 14.5°$.

12.7 (a) The spatial distribution of the electric field amplitude in the plane of the 5 light sources is

$$\begin{aligned} E(x,y) = E_0 \cdot [&\delta(x)\delta(y) + \delta(x - x_0)\delta(y) \\ &+ \delta(x + x_0)\delta(y) + \delta(x)\delta(y - y_0) \\ &+ \delta(x)\delta(y + y_0)]. \end{aligned}$$

The amplitude distribution in the diffraction plane is

$$\begin{aligned} E(x',y') &= A \cdot \iint E(x,y) e^{-2\pi i(v_x x + v_y y)} dx dy \\ &= AE_0\left[1 + e^{-2\pi i v_x x_0} + e^{2\pi i v_x x_0} + e^{-2\pi i v_y y_0} + e^{2\pi i v_y y_0}\right] \\ &= AE_0[1 + 2\cos(2\pi\, v_x x_0) + 2\cos(2\pi\, v_y y_0)] \\ &= 2AE_0[\cos^2(\pi\, v_x x_0) + \cos^2(\pi\, v_y y_0) - 3/2]. \end{aligned}$$

this corresponds to an amplitude cross-grid with maxima along the perpendicular lines x = const and y = const., superimposed by a constant background, which would disappear if the amplitude of the source at (0, 0) is four times as large as that of the other sources. The periods in the diffraction plane are $v_x = 2/x_0$ and $v_y = 2/y_0$. Since $v_x = \frac{x'}{\lambda f}$ and $v_y = \frac{y'}{\lambda f}$ the spatial separation between the lines of maximum intensity is $\Delta x' = \lambda \cdot f v_x = 2\lambda \cdot f/x_0$ and $\Delta y' = 2\lambda f/y_0$.

(b) If the sources at $(x_0, 0)$ and $(-x_0, 0)$ are extinguished, the term $\cos^2(\pi\, v_x, x_0)$ becomes zero and therefore the stripes parallel to the y-direction disappear. It remains a pattern of stripes parallel to the x-direction.

(c) If the source at (0, 0) is extinguished, the constant background disappears. Now one gets a pattern with the half distance $\Delta x' = \lambda f/x_0, \Delta y' = \lambda f/y_0$ between the stripes.

12.8 We regard a grating with grooves (width b and distance d) parallel to the y-axis. The field amplitude in the plane of the grating is then

$$\begin{aligned} E &= E_0 \quad \text{for} \quad md + b \le x \le (m+1)d \\ &= 0 \quad \text{for} \quad md < x < md + b. \end{aligned}$$

This can be written as the product

$$E(x,y) = E_0\,\text{rect}\left(\frac{x}{b}\right) * \sum_{m=1}^{N} \delta(x - m \cdot d) = E_1 * E_2,$$

where the rectangle function $\text{rect}(x/b) = 1$ for $0 \le x/b \le 1$ describes the constant field amplitude within the slit width b. The field amplitude in the diffraction plane is given by the Fourier-transform of $E(x,y)$ which can be written as the convolution of the one-dimensional functions $E_1(x)$ and $E_2(y)$ (see Sect. 10.8).
Since the slits extend in y-direction from $-\infty$ to $+\infty$ there is no interference structure in the y-direction. The Fourier-transform of the rectangle function gives in the focal plane of the lens with focal length f

$$\mathcal{F}_1(v_x) = \frac{\sin(\pi b v_x)}{\pi\, v_x} \quad \text{with} \quad v_x = \frac{x'}{\lambda \cdot z} \approx \frac{\alpha}{\lambda}$$

for $z = f$, while the Fourier-transform of the delta function is

$$\mathcal{F}(\delta(x - md)) = e^{-2\pi m d v_x}$$

Finally we get the field amplitude in the observation plane

$$\begin{aligned} E(x',y') &= E_0 \cdot \delta(v_y) \frac{\sin(\pi b v_x)}{\pi v_x} \sum_{m=1}^{N} e^{-i2\pi\, m d v_x} \\ &= E_0 \cdot \delta(v_y) b \frac{\sin(\pi b v_x)}{\pi\, b v_x} \\ &\quad \cdot e^{-i\pi d v_x (N+1)} \frac{\sin(\pi N d v_x)}{\sin(\pi d v_x)}. \end{aligned}$$

The intensity distribution of the Fraunhofer diffraction pattern is then

$$I(x',y') \propto |E(x',y')|^2 = E_0^2 \cdot \delta(v_y) b^2 \cdot \frac{\sin^2(\pi b v_x)}{(\pi b v_x)^2} \frac{\sin^2(\pi N d v_x)}{\sin^2(\pi d v_x)},$$

This had been already derived in Sect. 10.8 in another way.
For $b = d/2$ we can rewrite this with $\sin 2x = 2\sin x \cdot \cos x$ as

$$I(x',y') = I_0 \cdot \delta(v_y) b^2 \frac{\sin^2(\pi N b v_x)\cos^2(\pi N b v_y)}{(\pi b v_x)^2 \cos^2(\pi b v_x)}.$$

The corresponding intensity distribution is shown in Fig. 10.42.

12.9 The minimum acceptable difference Δn of the refractive indices is according to (12.49)

$$\Delta n = n_2 - n_1 > \frac{m_s^2 \lambda^2}{4a^2(n_2+n)} \Rightarrow n < \left[n_2^2 - \left(\frac{m_s \lambda}{2a}\right)^2 \right]^{1/2}.$$

(a) $(a) m_s = 1: \Rightarrow n < \left[4 - \left(\frac{0.6}{4}\right)^2\right]^{1/2} = 1.994 \Rightarrow \Delta n \geq 0.006.$

The parameter h, p and q are determined from the relations

$$h = \frac{2\pi}{\lambda}\sqrt{n_2^2 - n^2} \geq \frac{2\pi}{0.6}\sqrt{4 - 1.994^2}\,\mu\text{m}^{-1} = 1.621\,\mu\text{m}^{-1}.$$

$$h = \sqrt{n_2^2 k^2 - \beta^2} \Rightarrow \beta = \sqrt{n_2^2 k^2 - h^2} \leq k \cdot n = 20.88\,\mu\text{m}^{-1}$$
$$\Rightarrow \cos\vartheta = \frac{\beta}{n_2 k} \geq 0.997 \Rightarrow \vartheta \leq 4.44^\circ.$$

this mode can only propagate through the fiber, if the angle of the **k**-vector against the z-axis is below 4.44°. Its penetration depth h into the surrounding medium must be calculated from the coefficients p and q and is for $n_1 = n_3 = n$ and $p = q$ according to (12.45)

$$p = h^2/d = \frac{1.621^2}{2}\,\mu\text{m}^{-1} = 1.3\,\mu\text{m}^{-1}.$$

The amplitude of the guided wave has decayed in the surrounding medium for a penetration depth of 0.76 µm to 1/e of its value in the core of the fiber.

(b) $m_s = 2 :\Rightarrow n \leq \left(4 - \left(\frac{1.2}{4}\right)^2\right)^{1/2} = 1.227 \Rightarrow \Delta n \geq 0.023,$

$h \geq 3.17\,\mu\text{m}^{-1} \Rightarrow \vartheta \leq 8.7^\circ,$

$p \geq \frac{3.17^2}{2}\,\mu\text{m}^{-1} = 5.02\,\mu\text{m}^{-1}$. The penetration depth into the surrounding is now only 0.2 µm.

(c) $m_s = 3 :\Rightarrow n \leq \left[4 - \left(\frac{1.8}{4}\right)^2\right]^{1/2} = 1.949 \Rightarrow$

$\Delta n \geq 0.0051,\ h \geq k \cdot \sqrt{n_2^2 - n^2} \geq 4.70\,\mu\text{m}^{-1},$

$q = p \geq 11.0\,\mu\text{m}^{-1}; \vartheta \leq 13^\circ.$

Penetration depth = 0.09 µm.

12.10 The frequency width of the pulse is with $\Delta\nu \cdot \Delta t = 1 \Rightarrow \Delta\nu = 10^{12}\,\text{s}^{-1}$

with $\lambda = c/\nu \Rightarrow |\Delta\lambda| = c/\nu^2 \Delta\nu = (\lambda^2/c) \cdot \Delta\nu$
$\Rightarrow \Delta\lambda = 1.3^2 \times 10^{-12}/(3 \times 10^8) \times 10^{12}\,\text{m}$
$= 5.6 \times 10^{-9}\,\text{m} = 5.6\,\text{nm}.$

$$\Delta t = \frac{L}{c} \cdot \Delta n = \frac{L}{c} \cdot \frac{dn}{d\lambda} \cdot \Delta\lambda$$
$$\Rightarrow L = \frac{c \cdot \Delta t}{\frac{dn}{d\lambda} \cdot \Delta\lambda} = \frac{3 \times 10^8 \times 10^{-12}}{2 \times 10^{-6} \cdot 5.6}\,\text{m} = 26.8\,\text{m}.$$

After L = 26.8 m the width of the pulse has increased to 2 ps.

12.11 The light path through the gradient fiber is $r(z)$ where the z-axis is the symmetry axis of the fiber.
The differentiation $d\boldsymbol{r}/d\boldsymbol{s}$ in (12.60) changes to dr/dz and the index gradient is $\nabla n = (dn/dr) \cdot \hat{\boldsymbol{e}}_r$ ($\hat{\boldsymbol{e}}_r$ is the unit vector in the radial direction) because n does not depend on z.
From (12.60) it follows:

$$\frac{d}{ds}\left(n \cdot \frac{d\boldsymbol{r}}{ds}\right) \rightarrow n(r) \cdot \frac{d^2 r}{dz^2}\hat{\boldsymbol{e}}_r \Rightarrow \frac{d^2 r}{dz^2} = \frac{1}{n(r)} \cdot \frac{dn}{dr}.$$

12.12 The maximum angle α_0 appears when the light beam passes through the symmetry axis.

$$\text{From } r(z) = a \cdot \sin\left(\frac{\sqrt{2\Delta}}{a} \cdot z\right) \Rightarrow \frac{dr}{dz} = \sqrt{2\Delta} \cdot \cos\left(\frac{\sqrt{2\Delta}}{a} \cdot z\right).$$

For $r = 0$ is $\frac{\sqrt{2\Delta a}}{a} \cdot z = n \cdot \pi$

$$\Rightarrow z(r=0) = \frac{n \cdot a \cdot \pi}{\sqrt{2\Delta}}$$
$$\Rightarrow \left.\frac{dr}{dz}\right|_{r=0} = \sqrt{2\Delta} \cdot \cos(n \cdot \pi) = \sqrt{2\Delta} = \tan\alpha_0.$$

Values of the Physical Fundamental Constants[a]

Quantity	Symbol	Value	Unit	Relative uncertainty in 10^{-6}
Speed of light in vacuum	c	29,9792,458	m s^{-1}	exact
Gravitation constant	G	$6.6730 \cdot 10^{-11}$	m^3 kg^{-1} s^{-2}	22
Planck constant	h	$6.62607015 \cdot 10^{-34}$	J s	exact
Reduced Planck constant	$\hbar$	$1.054571817\ldots \cdot 10^{-34}$	J s	exact
Molar gas constant	R	8.314462618	J mol^{-1} K^{-1}	exact
Avogadro constant	N_A	$6.02214076 \cdot 10^{23}$	mol^{-1}	exact
Lohschmidt constant (T = 273.15 K, p = 100 kPa)	N_L	$2.6516467 \cdot 10^{25}$	m^{-3}	0.57
Boltzmann constant R/N_A	k	$1.380649 \cdot 10^{-23}$	J K^{-1}	exact
Molar volume (T = 273.15 K, p = 101,325 Pa)	V_M	$22.41396454 \cdot 10^{-3}$	m^3 mol^{-1}	exact
(T = 273.15 K, p = 100 kPa)		$2271095464 \cdot 10^{-3}$	m^3 mol^{-1}	exact
Elementary charge	e	$1.602176634 \cdot 10^{-19}$	As $\stackrel{\text{Def}}{=}$ C	0.0003
Electron mass	m_e	$9.1093837015 \cdot 10^{-31}$	kg	0.0003
Proton mass	m_p	$1.67262192369 \cdot 10^{-27}$	kg	0.0003
Magnetic constant	$m\mu_0$	$1.256637062 \cdot 10^{-6}$	V s A^{-1} m^{-1}	0.00015
Electric constant $1/(\mu_0 c^2)$	ε	$8.8541878128\ldots \cdot 10^{-12}$	A s V^{-1} m^{-1}	0.00015
Fine structure constant $\mu_0 c e^2/2h$	α	$7.2973525693 \cdot 10^{-3}$	–	0.00015
Rydberg constant $m_e c\alpha^2/2h$	Ry_∞	$1.0973731568160 \cdot 10^{7}$	m^{-1}	0.0000019
Bohr radius $\alpha/(4\pi Ry_\infty)$	a_0	$5.29177210903 \cdot 10^{-11}$	m	0.00015
Proton–Electron mass ratio	m_p/m_e	1836.15267343	–	0.00006
Electron charge-to-mass quotient	$-e/m_e$	$-1.75882001076 \cdot 10^{11}$	C kg^{-1}	0.0003
Proton charge-to-mass quotient	$+e/m_p$	$+9.57882001560 \cdot 10^{7}$	C kg^{-1}	0.00031
Atomic mass unit $\frac{1}{12}m(^{12}C)$	AMU	$1.66053906660 \cdot 10^{-27}$	kg	0.0003

Conversion factor
1 eV = $1.60217653 \cdot 10^{-19}$ J
1 eV/hc = 8065.541 cm^{-1}
1 Hartree = 27.2113845 eV
1 Hartree/hc = $2.194746313 \cdot 10^{5}$ cm^{-1}
[a]CODATA, international recommended values (NIST 2018)

Astronomical Constants

Mass of Earth	$M_E = 5.9736 \times 10^{24}$ kg	
Mass of Moon	$M_M = 7.35 \times 10^{22}$ kg	$\triangleq 0.0123\ M_E$
Mass of sun	$M_0 = 1.989 \times 10^{30}$kg	$\triangleq 3.33 \times 10^{5} M_E$
Radius of sun	6.96×10^{8} m	
Distance earth-moon		
Minimum (Perihel)	3.564×10^{8} m	
Maximum (Aphel)	4.067×10^{8} m	
Mean distance earth-sun	1.496×10^{11} m	
1AU = astronomical unit	$1.49597870700 \times 10^{11}$ m	

Useful Conversion Factors

Lengths		
1 Å	1 Ångström	$\triangleq 10^{-10}$ m $\triangleq$ 100 pm
1 f	1 Fermi	$\triangleq 10^{-15}$ m $\triangleq$ 1 fm
1 AE	1 Astronomical unit AU	$\triangleq 1.49598 \times 10^{11}$ m
1 ly	1 light year	$\triangleq 9.46 \times 10^{15}$ m
1 pc	1 Parsec	$\triangleq 3.09 \times 10^{16}$ m

Time
1 Year = 3.156×10^{7} s
1 Day = 8.64×10^{4} s

Energy	
1 eV	1.60218×10^{-19} J
1 kWh	3.6×10^{6} J
1 kcal	4.1868 kJ
1 kcal/mol	4.34×10^{-2} eV pro Molekül
1 kJ/mol	1.04×10^{-2} eV pro Molekül

From $E = mc^2$ is follows: 1 kg $\cdot$ c^2 = 8,98755 $\times$ 10^{16} J
With k = 1,380658 $\times$ 10^{-23} J K^{-1} follows for 1 eV $\triangleq$ $k \cdot T$ bei T = 11604 K
With $h \cdot \nu = E$ follows for the frequency ν of electromagnetic radiation $\nu = E/h = 2.418 \times 10^{14}$ Hz eV^{-1}

Angles	
1 rad	57.2958°
1°	0.0174 rad
1′	2.9×10^{-4} rad
1″	4.8×10^{-6} rad

Mathematical constants	
π	3.141592653589
e	2.718281828459
ln2	0.693147180559
$\sqrt{2}$	1.414213562373
$\sqrt{3}$	1.732050807568

Approximation formulas for $\lvert x\rvert \ll 1$	
$(1 \pm x)^n \approx 1 \pm nx$	$\cos x \approx 1 - x^2/2$
$\sqrt{1 \pm x} \approx 1 \pm \frac{1}{2}x$	$e^x \approx 1 + x$
$\sin x \approx x$	$\ln(1 + x) \approx x$

The Greek Alphabet

Letters	Name	Letters	Names
A, α	Alpha	N, ν	Ny
B, β	Beta	Ξ, ξ	Xi
Γ, γ	Gamma	O, o	Omikron
Δ, δ	Delta	Π, π	Pi
E, ε	Epsilon	P, ϱ	Rho
Z, ζ	Zeta	$\sum$, σ	Sigma
H, η	Eta	T, τ	Tau
Θ, ϑ	Theta	Υ, υ	Ypsilon
I, ι	Jota	ϕ, φ	Phi
K, κ	Kappa	X, χ	Chi
Λ, λ	Lambda	Ψ, ψ	Psi
M, μ	My	Ω, ω	Omega

Index

W. Demtröder, *Electrodynamics and Optics*, Undergraduate Lecture Notes in Physics, https://doi.org/10.1007/978-3-030-02291-4